Expert Systems With Applications 112 (2018) 34–60

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Latent association rule cluster based model to extract topics for classification and recommendation applications

Fabiano Fernandes dos Santos[a], Marcos Aurélio Domingues[b,*], Camila Vaccari Sundermann[a], Veronica Oliveira de Carvalho[c], Maria Fernanda Moura[d], Solange Oliveira Rezende[a]

[a] Institute of Mathematics and Computer Science - University of São Paulo Avenida Trabalhador São Carlense, 400, São Carlos, SP, 13566-590, Brazil
[b] Department of Informatics - State University of Maringá Avenida Colombo, Maringá, PR, 5790, 87020-900, Brazil
[c] Institute of Geosciences and Exact Sciences - State University of São Paulo Avenida 24 A, Rio Claro, SP, 1515, 13506-900, Brazil
[d] Embrapa Agricultural Informatics Avenida Dr. André Tosello, Campinas, SP, 209, 13083-886, Brazil

ARTICLE INFO

Article history:
Received 19 August 2017
Revised 7 June 2018
Accepted 8 June 2018
Available online 15 June 2018

Keywords:
Document representation
Topic model
Association rules
Clustering
Text classification
Context-aware recommender systems

ABSTRACT

The quality of any text mining technique is highly dependent on the features that are used to represent the document collection. A classical form of document representation is the vector space model (VSM), according to which the documents are represented as vectors of weights that correspond to the features of the documents. The bag-of-words model is the most popular VSM approach due to its simplicity and general applicability, but this model does not include term dependency and has a high dimensionality. In the literature, several models for document representation have been proposed in order to capture the dependency of terms. Among them, the topic model representation is one of the most interesting approaches - since it describes the collection of documents in a way that reveals their internal structure and the interrelationships therein, and also provides a dimensionality reduction. However, even for topic models, the efficient extraction of information concerning the relations among terms for document representation is still a major research challenge. In order to address this issue, we proposed thelatent association rule cluster based model (LARCM). The LARCM is a non-probabilistic topic model that makes use of association rule clustering to build a document representation with low dimensionality in such a way that each feature (i.e., topic) is comprised of information concerning relations among the terms. We evaluated the interpretability of the topics obtained by using our proposed model against the ones provided by the traditional latent dirichlet allocation (LDA) model and the LDA model using a document representation that includes correlated terms (i.e., bag-of-related-words). The experimental results indicated that the LARCM provides topics with better interpretability than the LDA models. Additionally, we used the topics obtained by the LARCM in two different applications: text classification and page recommendation. With respect to text classification, the topics were used to improve document collection representation. Concerning page recommendation, topics were used as contextual information in context-aware recommender systems. Results have shown that the topics provided by the LARCM can be used to improve both applications.

1. Introduction

The text mining research field has been the object of great attention in recent years due to the large amount of available text data. Nowadays, unstructured textual data is commonly created in any application scenario (Aggarwal & Zhai, 2012). Due the nature of this data, effective information retrieval is becoming more and more difficult without efficient organization, summarization and indexing of document contents (Liu, 2011).

The quality of the results provided by any text mining technique is highly dependent on the features that are used to represent the document collection (Aggarwal & Zhai, 2012; Keikha, Khonsari, & Orumchian, 2009; Shafiei et al., 2007). A classical document representation is the vector space model (VSM). In this model, each document is represented as a vector of weights that correspond to the features of the document. The most popular

* Corresponding author.
E-mail addresses: fabianof@icmc.usp.br (F.F. dos Santos), madomingues@uem.br (M.A. Domingues), camilavs@icmc.usp.br (C.V. Sundermann), veronica@rc.unesp.br (V.O. de Carvalho), maria-fernanda.moura@embrapa.br (M.F. Moura), solange@icmc.usp.br (S.O. Rezende).

https://doi.org/10.1016/j.eswa.2018.06.021
0957-4174/

VSM is the bag-of-words, which is a model that enables the easy utilization of several text mining techniques (Salton, 1989).

VSM models have become very effective due to their simplicity and general applicability. However, these models have high dimensionality and do not include term dependencies (Aggarwal & Zhai, 2012). High dimensionality has always been a great challenge for all machine learning algorithms, and the "*curse of dimensionality*" has been studied for a long time (Shafiei et al., 2007). Additionally, the VSM models do not include term dependencies and because of that, they cannot distinguish between the different meanings of a polysemous word in different contexts, or realize the common meaning between synonyms. It also fails to recognize multi-word expressions (e.g., "*artificial intelligence*") (Cheng, Miao, Wang, & Cao, 2013; Farahat & Kamel, 2011; Kalogeratos & Likas, 2012).

There are many evidences in the literature that the term correlation improves the results of text mining tasks such as text classification (Figueiredo et al., 2011; Keikha et al., 2009; Le & Mikolov, 2014), text clustering (Beil, Ester, & Xu, 2002; Cheng et al., 2013; Kalogeratos & Likas, 2012; Marcacini, Correa, & Rezende, 2012; Rossi & Rezende, 2011; Zhang, Yoshida, Tang, & Wang, 2010), topic extraction (Lau, Baldwin, & Newman, 2013; Wallach, 2006; Wu, Lee, & Wang, 2010; Zhu, Fukazawa, Karapetsas, & Ota, 2012) and textual content recommendation (Domingues et al., 2015; Domingues, Sundermann, Manzato, Marcacini, & Rezende, 2014; Manzato, Domingues, Marcacini, & Rezende, 2014). However, the efficient extraction of correlated terms for document representation is still a major research challenge (Cheng et al., 2013; Figueiredo et al., 2011; Gao, Xu, Li, & Liu, 2013).

The document representation models that make use of correlated terms usually face a great challenge in keeping a good trade-off among (i) number of extracted features, (ii) computational performance and (iii) interpretability of new features. The items (i) and (ii) are closely related, since the number of possible feature combinations for a collection containing n terms is 2^n. Therefore, there is a computational cost for extracting features from large collections of text, besides the challenge of identifying which combination is significant for the application (Figueiredo et al., 2011). With respect to the interpretability of new features in item (iii), Chang, Boyd-Graber, Gerrish, Wang, and Blei (2009) discuss a divergence between the objective and subjective evaluation of topic models. The authors concluded that models with good objective evaluation may not provide topics (new features) with good interpretability. In Section 2, we present works proposed in the literature and discuss them in terms of these challenges.

In the literature, topic models are some of the most interesting approaches for handling term correlation in document representation. This approach provides a strategy for dimensionality reduction by creating new features that represent the main topics or ideas from the document collection. In addition, topic organization is able to cluster terms with the same meaning in groups and place a term in more than one group, if it has multiple meanings in the text collection. A clustering of association rules can be used to implement a topic model approach that makes use of correlated terms, avoiding the previously discussed issues (Liu, 2011; Pôssas, Ziviani, Meira, & Ribeiro-Neto, 2002; Rossi & Rezende, 2011). Association rules represent item (a term in this paper) correlations or co-occurrences (Agrawal & Srikant, 1994). According to Liu (2011), there are two main advantages: 1) association rules mining algorithms are very efficient, and 2) association rules are easy to understand. In addition, the clustering task provides a dimensionality reduction strategy.

In this work, we propose the latent association rule cluster based model (LARCM). This is a new non-probabilistic topic modeling approach that makes use of the clustering of association rules as to provide a model for document representation with correlation of terms and low dimensionality. In our proposal, association rules are built for each document in order to extract the correlated terms. We call these relations among terms the *local context* of relations. Then, we apply a clustering process to all association rules in order to discover the *general context* of relations. Each cluster becomes a feature (i.e., a topic) for the new document representation. The main idea behind our proposal is that the neighborhood of correlated terms will reveal (a) different terms used in the same context or with the same meaning, and (b) identical terms used in different contexts or with different meanings.

We evaluated the interpretability of the topics obtained with the LARCM against the ones provided by the traditional latent dirichlet allocation (LDA) model and the LDA model using document representation that includes correlated terms (i.e., bag-of-related-words). The experimental results demonstrated that LARCM provides topics with better interpretability than those produced by the LDA models. Additionally, we used the topics obtained with the LARCM in two different applications: text classification and page recommendation. In text classification, the topics were used to improve document collection representation. Concerning page recommendation, the topics were used as contextual information in context-aware recommender systems. The results have shown that the topics provided by the LARCM can be used to improve both applications.

Thus, our main contributions in this work can be summarized as follows:

- LARCM: a new non-probabilistic topic modeling approach that provides a document representation with correlation of terms and low dimensionality;
- Utilization of the LARCM to improve two text mining tasks, i.e., text classification and page recommendation.

The rest of the paper is structured as follows: Section 2 discusses a few topic modeling approaches that make use of correlated terms. Our proposal, which is a non-probabilistic topic modeling approach, is described in Section 3. In Section 4 we evaluate our model, while in Section 5 we use our proposal to improve text classification and page recommendation. Finally, in Section 6, we present our conclusion as well as future directions.

2. Related work

As already stated, document representation models that make use of correlated terms can improve the results of many text mining tasks. In this section, we discuss the main such representation models that make use of correlated terms. In order to provide a good overview of the field, we have decided to present methods that produce a term-based representation as output, and methods that produce a topic-based representation. In this work we are interested in the latter and, mainly, in the representations that produce a latent dimension.

Generalized vector space model (GVSM) (Wong, Ziarko, & Wong, 1985) is one of the first models that incorporate correlation of terms in document representation. The authors proposed an alternative interpretation to the vector space model (VSM) in which the vectors of the indexing terms are comprised of smaller components, called minterms, which are derived from the text collection. The model adopts as basic principle the idea that the co-occurrence of terms within the documents induces the dependency among these terms (Baeza-Yates & Ribeiro-Neto, 2011; Pôssas et al., 2002). However, according to Baeza-Yates and Ribeiro-Neto (2011), it is not clear in which situations the GVSM outperforms the classical VSM model. In addition, GVSM can be quite complex and computationally expensive for large collections, given that the number of minterms that need to be computed usually equals the number of documents in the collection. Thus, the main contribution of

the proposal is the establishment of a formal framework in which the dependencies among terms can be adequately represented, introducing new ideas that are theoretically important (Baeza-Yates & Ribeiro-Neto, 2011; Pôssas et al., 2002). On the other hand, several extensions of the GVSM, which aim to compensate for some of the shortcomings of the original proposal, have been proposed in the literature (Billhardt, Borrajo, & Maojo, 2002; Cheng et al., 2013; Farahat & Kamel, 2011; Kalogeratos & Likas, 2012).

The models based on phrases or compound terms, such as those presented in (Badawi & Altinçay, 2014; Figueiredo et al., 2011; Pôssas et al., 2002; Rossi & Rezende, 2011), aim to construct new attributes comprised of 2 or more terms, introducing a concept different from the one dealing with individual terms. The new attributes can be added to the bag-of-words representation or can override the original attributes. According to Rossi and Rezende (2011), the models proposed in the literature which are based on phrases or compound terms are based on n-grams and sets of words. With the n-grams model, attributes are constructed by sequences of n adjacent terms that appear in the text collection. In that model based on a set of words, attributes are also comprised of n terms that appear in the text collection but do not need to be adjacent. In general, the works in the literature for this type of model propose the execution of a co-occurrence analysis considering the whole document collection. For example, the model proposed in Figueiredo et al. (2011) makes use of the co-occurrence of terms in documents to construct attributes called c-features (compound-features). In order to obtain the most significant attributes, the model uses the label of each document in the training set. The authors aimed to obtain these attributes with a low-cost computational strategy, since the number of possible combinations involving the original attributes for a collection with n terms is 2^n. In Badawi and Altinçay (2014) a similar framework is proposed based on the co-occurrence statistics of pairs of terms in which each termset (compound term) is assigned a nonzero weight if either or both of the terms occur in the document. In their work, Rossi and Rezende (2011) introduced the bag-of-related-words model, which uses association rules to extract, from the documents, indexing terms comprised of strongly correlated simple words, and simple words that often occur in the collection. The indexing terms are utilized to construct the representation for the text collection. The main advantages of this approach are: (1) the dimensionality of the representation is not significantly increased, as in most of the models proposed in the literature, (2) there is an automatic process to generate significant correlated terms, (3) labeled collections are not necessary, and (4) the entire collection does not need to be scanned before term selection is made, because the model is applied to each document individually. Finally, to make the phrases or compound terms models viable for practical applications, it is necessary to reduce the dimensionality of representations by applying an attribute selection approach.

Regarding the other correlated terms representations found in the literature, the ones based on the extraction of latent features, such as latent semantic analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), and the ones based on topic extraction (Blei, Ng, & Jordan, 2003; Hofmann, 1999), stand out for the quality of the results and the good interpretability of the extracted features (i.e., topics). These methods can also be described as non-probabilistic and probabilistic approaches. The latent dirichlet allocation (LDA) (Blei et al., 2003) model is one of the most prominent techniques for topic extraction, since it treats each document as a mixture of topics, and each topic as a mixture of words. Thus, the documents can "overlap" in terms of content, rather than being separated into discrete groups. In Blei et al. (2003), a generative model is proposed to describe a collection of documents from a reduced set of topics. This model becomes attractive to find groups of terms that appear frequently together in the documents (Hu, Boyd-Graber, Satinoff, & Smith, 2014; Zhu et al., 2012). Also, topic model approaches provide good results when used in many important tasks, even in scenarios where non-textual features combined with textual data are used, as presented in Zhang, Yuan, Lian, Xie, and Ma (2016), Xie et al. (2016), Zoghbi, Vulić, and Moens (2016), Kefalas and Manolopoulos (2017), Song, Zhang, Duan, Hossain, and Rahman (2017), Guo, Barnes, and Jia (2017), Tapi, Bringay, Lavergne, Mollevi, and Opitz (2017), Chen and Ren (2017), Momtazi (2018) and Chien, Lee, and Tan (2018).

With respect to the use of term correlation in topic models, the proposal presented in Wallach (2006) is one of the first to incorporate correlated terms in LDA to extract topics. Wallach (2006) used dirichlet's hierarchical language models to extend the original LDA algorithm in order to take into account - during the inference of the topics - the weight of a term, conditionally evaluated with respect to the weight of the previous term. In Lau et al. (2013), the authors claim that, although the results obtained with proposed methods based on modifications of the inference algorithm are theoretically interesting, the inference algorithm becomes computationally more complex, and this can make its usage impossible for the end user. Their conclusions are based on the evaluation of the utilization of correlated terms in the topic model obtained by applying the classic LDA model. More recent proposals include frequent itemset extraction techniques in order to make use of the dependency of terms in the extraction of topics without increasing the complexity of the process (Gao et al., 2013; Kim, Park, Lu, & Zhai, 2012; Zhu et al., 2012). However, the quality of the topics obtained with these proposed methods has not been thoroughly evaluated.

In Table 1 we summarize the models discussed in this section which make use of term correlation. We relate each model to the issues described in Section 1, which the model intends to address.

Term dependency can be made use of at different levels in the document collection. The most common case is the utilization of correlated terms, that is, two or more terms with different individual meanings which produce a new meaning when they occur in close proximity. For example, the term "artificial intelligence" has a specific meaning whereas the terms "intelligence" and "artificial", utilized separately, have other meanings. The correlated terms create a dependency among terms in a local context, because the meaning of terms depends only on their proximity in the document, regardless of the behavior of the rest of the collection.

Another type of dependency found in the literature is the one utilized by topic extraction models. In these models, each topic is defined by a set of simple terms used to describe a particular subject. In this case, the dependency among the terms is due to its use in a shared context. Thus, the meaning of a term will be based on the context, and will not depend on the proximity to other terms. This type of term dependency occurs in a more general context, since the meanings of the terms depend on the topics in the collection, which are not known during the construction of the model.

Correlated terms may appear to be strongly related in a topic. However, there is no guarantee that the occurrence of these terms in a shared topic is related to their dependency in a local context. For example, a topic described by the terms "intelligence" and "artificial" can be found in both computer and psychology documents referring to completely different subjects. However, the term "artificial intelligence" will be more likely referring to computer documents.

Motivated by the challenges previously discussed in this work as well as the results obtained through the models found in the literature which combine strategies of topic extraction and dependency of terms, in this paper we introduce the LARCM, a new non-probabilistic topic model that makes use of the dependency among terms in order to represent collections of documents and provide a

Table 1
Models that make use of term correlation and the issues that they propose to address.

Model	(i) Number of features	(ii) Computational effort	(iii) Interpretability of the new features
Based on the vector space model			
Context Vector Model- CVM-VSM (Billhardt et al., 2002)			X
Generalized Vector Space Model- GVSM (Wong et al., 1985)			X
Semantic Similarity Based on Term-Term Correlations (Farahat & Kamel, 2011)			X
Global Term Context Vector-VSM- GTCV-VSM (Kalogeratos & Likas, 2012)			X
Coupled Term-Term Relation Model- CRM (Cheng et al., 2013)			X
Based on phrases or compound terms			
Set-Based Model (Pôssas et al., 2002)		X	X
C-Features (Figueiredo et al., 2011)		X	X
Bag-of-Related-Words (Rossi & Rezende, 2011)		X	X
Termsets (Badawi & Altinçay, 2014)		X	X
Based on latent features			
LSI/LSA (Deerwester et al., 1990)	X		X
pLSA (Hofmann, 1999)	X		X
LDA (Blei et al., 2003)	X		X
Frequent Pattern-Based Data Enrichment Approach (Kim et al., 2012)	X		X
Word-Pair Latent Dirichlet Allocation Model - wpLDA (Zhu et al., 2012)	X		X
Two-Stage Approach Topic Model (Gao et al., 2013)	X		X

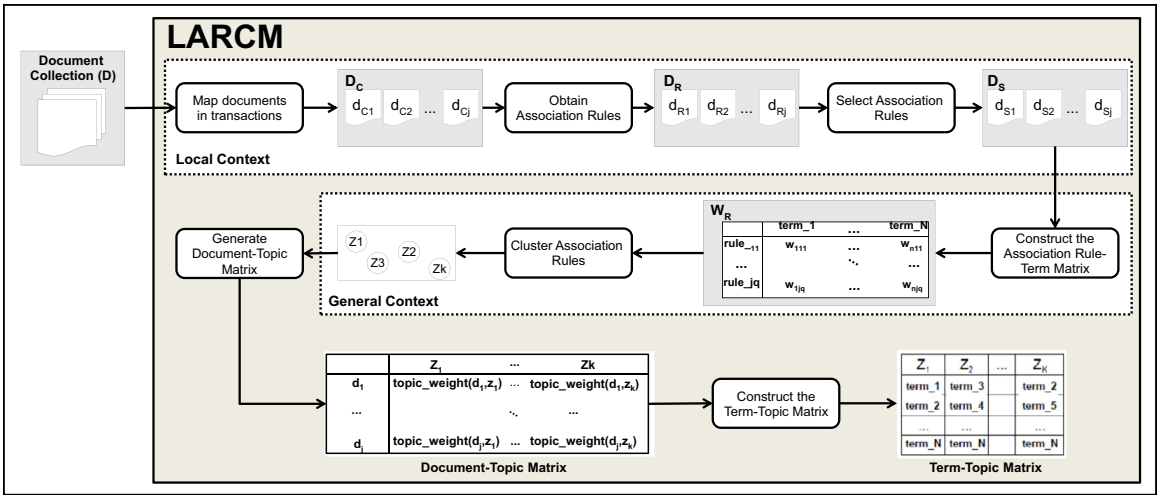

Fig. 1. Latent association rule cluster based model (LARCM).

representation with both high interpretability and reduced dimensionality.

3. LARCM: latent association rule cluster based model

In this section we will present our LARCM, which is a representation model that makes use of the relations among terms in both local and general contexts, identifying topics in document collections. In this model, the local context of the relation among terms is defined through the identification of their co-occurrences, using information extracted from the document in which they occur, i.e., it is independent from the occurrence of the terms in other documents within the collection. This information is utilized in this work for the identification of correlated terms, such as “data mining” or “artificial intelligence”. For the general context, we identify some relations among the terms extracted from the local context. To obtain the topics, the relations among terms are clustered so that each cluster will contain the terms strongly related to a topic expressed in the document collection. Thus, we address important problems such as polysemy, which is still a major challenge when we have correlated terms. For example, the terms “data mining” and “knowledge extraction” can be used in documents with similar meaning, even though they are represented by different terms. The LARCM is illustrated in Fig. 1.

In Fig. 1, the local context captures the correlation among the terms in each document through the co-occurrence obtained with the extraction of association rules. To capture the general context, we propose an intermediate representation for the association rules that allows them to be represented in the original vector space of the documents. Each association rule is represented by a vector of distinct terms which appear in the collection of documents, and the weight of each term is given by its frequency in the transactions covered by the association rule. Thus, we can apply similarity measures, such as the cosine measure, to determine the general context of each relation while taking into account all other relations of the collection. In this way, the information on the neighborhood of terms, obtained by analyzing the set of transactions covered by each rule, helps to identify different terms used in a same context or with a same meaning, and also identical terms which are used in different contexts or with different meanings. For example, the term “cluster” may refer to a data mining technique when it appears close to the term “algorithm”, or it may refer to a group of computers when it appears close to the term “network”. Note that the LARCM is applied to a collection of documents that is pre-processed with normalization, stopword removal, cleaning and standardization of documents, as described in Nogueira et al. (2008).

Table 2
Notation used to describe the LARCM.

Notation	Description
D	Document collection.
d_j	A document in the collection D.
T	Set of distinct terms in the collection D.
w_{ij}	The weight of a document d_j with respect to its term $t_i \in T$.
W	Document-term matrix to represent the document collection (i.e., documents x terms).
D_C	A collection containing the documents of D mapped into transactions.
d_{Cj}	A set of transactions obtained by mapping the document $d_j \in D$.
A	A set of terms $t_i \in T$ for a document d_j such that $w_{ij} > 0$.
D_R	A collection containing the association rules obtained for each set of transactions d_{Cj}.
d_{Rj}	A set of association rules obtained from the set of transactions d_{Cj}.
r_{jq}	An association rule.
D_S	A collection of sets of association rules selected to represent the local context.
d_{Sj}	A set of association rules selected to represent the local context.
W_R	An association rule-term matrix to represent the selected rules and the terms which are contained in such rules.
Z	A set of clusters generated from the W_R matrix.
size_window	The size of the window/sequence of words that will be analyzed.
supmin	Minimum support value to generate association rules.
confmin	Minimum confidence value to generate association rules.

To describe the LARCM in details, in the next subsections, we will consider the notation presented in Table 2. Let $D = \{d_1, d_2, \ldots, d_m\}$ be a collection with m documents, and let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n distinct terms in the collection. Each document d_j is represented by a vector of terms $\vec{d_j} = \{w_{1j}, w_{2j}, \ldots, w_{nj}\}$, where each weight w_{ij} quantifies the importance of the term $t_i \in T$ for the document $d_j \in D$. For the terms of the collection that are not presented in the document, $w_{ij} = 0$. The collection of documents is then represented by the matrix W with dimensions $m \times n$, known in the literature as the document-term matrix. Each line of W corresponds to a document in D, i.e., the vector $\vec{d_j}$, and each column describes the distribution of each term in the document collection.

3.1. Identifying local context of the relation

To obtain the local context for the relation among terms, we propose a process adapted from the bag-of-related-words model presented in Rossi and Rezende (2011). The bag-of-related-words model offers an efficient process for extracting correlated terms by using association rules. In this model, the association rules are extracted for each document and they are used to construct the correlated terms for the collection. As each document is processed independently, the correlated terms are obtained while taking into account the local context of the relation among terms. The bag-of-related-words model presents four steps (Rossi & Rezende, 2011): (1) mapping the textual document into transactions, (2) extracting association rules from the transactions, (3) using the itemsets of the rules to compound the features, and (4) using the features to construct the document-term matrix. From these four steps, we adapted the steps (1) and (2) for the LARCM. These two steps are adapted so that the information provided at the end of each step can be used to construct a matrix from which the general context for the relation among terms is obtained.

For the LARCM, we have adapted the document mapping so that it uses sliding windows. During the mapping, the first transaction contains only the first word of the document, the second contains the first two words, and so on, until we arrive to a transaction that contains a number of words equal to the window's size, which is defined by the parameter "*size_window*". After that, the window slides a word ahead, and considers the next "*size_window*" words of the document as a transaction. This sliding process is repeated until the end of the document is reached. The result of this process is a collection $D_C = \{d_{C1}, d_{C2}, \ldots, d_{Cm}\}$, where d_{Cj} corresponds to the transactions obtained by mapping the document $d_j \in D$. This mapping is illustrated in Fig. 2 and can be performed by using the Algorithm 1.

Algorithm 1 Mapping documents to transactions.

Require: Document collection D; Size of the window (*size_window*) defined for mapping the transaction.
Ensure: The collection D_C containing the documents of D mapped into transactions.

1: $D_C \leftarrow \varnothing$
2: **for** each document $d_j \in D$ **do**
3: $d_{Cj} \leftarrow \varnothing$
4: $transaction \leftarrow \varnothing$
5: $w \leftarrow 1$
6: **while** $|transaction| <= size\_window$ **and** $w <= |words\ in\ d_j|$ **do**
7: $transaction \leftarrow transaction \cup word_w$
8: $w \leftarrow w + 1$
9: $d_{Cj} \leftarrow d_{Cj} \cup transaction$
10: **end while**
11: **for** i $\leftarrow size\_window$ to $|words\ in\ d_j|$ **do**
12: $transaction \leftarrow \varnothing$
13: $w \leftarrow 0$
14: **while** $|transaction| <= size\_window$ **and** $i + w <= |words\ in\ d_j|$ **do**
15: $transaction \leftarrow transaction \cup word_{i+w}$
16: $w \leftarrow w + 1$
17: **end while**
18: $d_{Cj} \leftarrow d_{Cj} \cup transaction$
19: **end for**
20: $D_C \leftarrow D_C \cup d_{Cj}$
21: **end for**
22: Return D_C

In Fig. 2, each document is processed individually. The window is positioned at the beginning of the document (Fig. 2-a), and the words inside the window are added to the set of transactions until a transaction of size 4 is formed ($size\_window = 4$). Then, the window is shifted to the right by one word, and the following sequence of 4 words is added to the set of transactions (Fig. 2-b). This process is repeated until the end of the document is reached (Fig. 2-c), decreasing the window's size until we have only one word in it (Fig. 2-d). After that, the next document is processed in the same way. The process ends when all documents are processed.

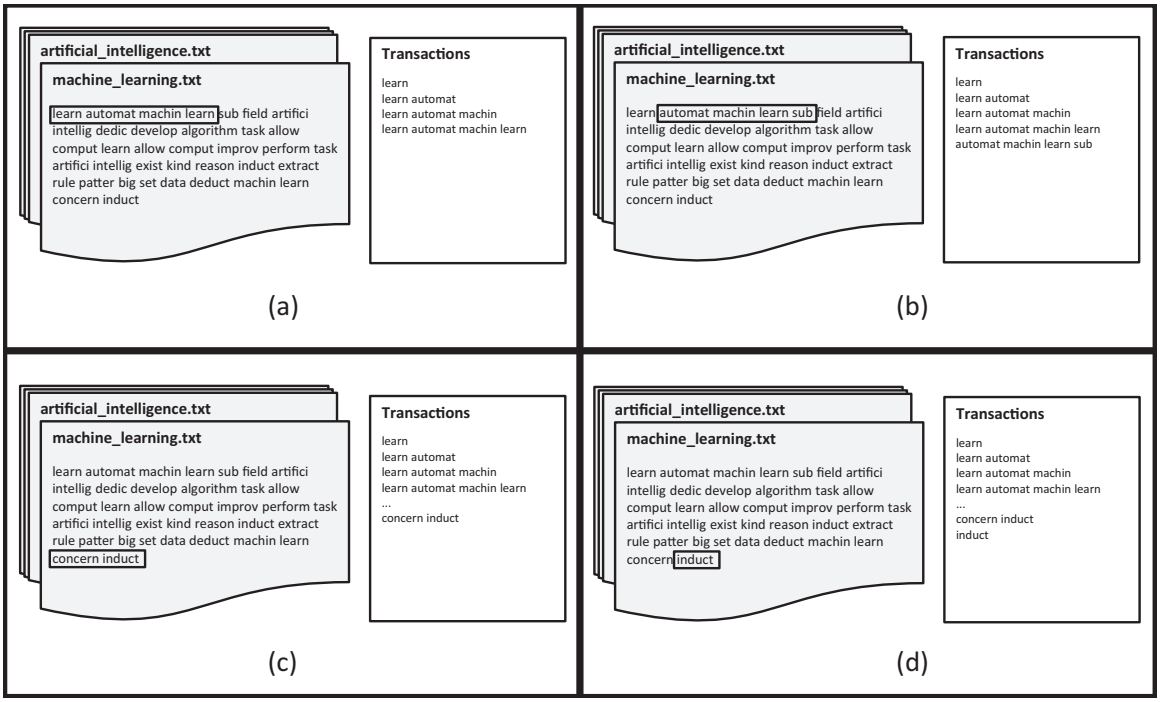

Fig. 2. Mapping of a document to a set of transactions.

The mapping of documents to transactions can be performed by using the Algorithm 1. In the algorithm, the pre-processed documents are traversed (line 2) and, for each document d_j, we obtain the words following the sequence in which they appear in the document (lines 6 to 10). By following the sequence, we form transactions from one word to *size_window* words. From the word in the *size_window* position, we add the consecutive words to a set that will form a new transaction (lines 11 to 19). We add the $i+w-th$ word to the set that will form a new transaction of size *size_window* (line 15). This process is repeated until the set has "*size_window*" words or until it is not possible to get more words from the document. Then, the transaction is added to the set of transactions d_{Cj} related to the document d_j (line 18). Finally, the set of transactions is added to the collection D_C (line 20). The algorithm repeats the process until all documents are mapped into transactions.

The next step to obtain the local context for the relation consists in applying the Algorithm 2 on the set D_C returned by the previous algorithm. In Algorithm 2, each set of transactions $d_{Cj} \in D_C$ (line 3) is processed by an association rules generation algorithm like Apriori (Agrawal & Srikant, 1994) (line 5). The minimum support (*supmin*) and minimum confidence (*confmin*) values are defined for each set of transactions d_{Cj}. In this work we used the value of $confmin = 0$ (line 2) in order to generate all possible association rules, since the rules that will be made use of throughout the process will be selected through the utilization of an objective measure in the next step. The *supmin* is calculated automatically for each document (line 4), as proposed by Rossi and Rezende (2011):

$$supmin(d_{Cj}) = \frac{(\sum_{\forall t_i \in A} w_{ij} \times size\_window)/|A|}{|d_{Cj}|} \times \alpha, \quad (1)$$

where A is the set of terms $t_i \in T$ for the document d_j such that $w_{ij} > 0$, $|A|$ is the total number of terms of d_j, *size_window* is the window size value defined for the mapping of transactions, $|d_{Cj}|$ is the total number of transactions of the document d_j and α is a factor that allows for the smoothing or intensifying of the threshold generated by the formula, and is usually defined as 1. With the Eq. (1), the user does not need to define a minimum support value. According to Rossi and Rezende (2011), this equation provides results comparable to those obtained with the manual setting of the *supmin* value.

In addition to composed terms, it is also important to have simple terms, since many of them have their own meanings. For this reason, we also considered generating association rules like $\varnothing \Rightarrow attribute$, i.e., rules containing only one item. These association rules will represent the simple terms. As a result, the Algorithm 2 returns a collection $D_R = \{d_{R1}, d_{R2}, \ldots, d_{Rm}\}$, where d_{Rj} corresponds to association rules obtained for each set of transactions d_{Cj}. The algorithm is illustrated in Fig. 3.

Finally, we use the Algorithm 3 to select the association rules that will represent the local context. For each set of association rules $d_{Rj} \in D_R$ (line 2), we compute the value of an objective measure in order to select association rules (line 3). The choice of an objective measure is very important, since each measure has its own semantics and influence on the type of correlated term that will be obtained by the algorithm. A set of objective measures (Carvalho, 2007) is presented in Table 3.

Algorithm 2 Generating association rules.

Require: Document collection D; Set of distinct terms for a collection T; Document-Term Matrix W; Size of the window (*size_window*) defined for mapping the transaction; Factor α; Collection of documents mapped into transactions D_C.
Ensure: The collection of association rules D_R for the documents in D_C.
1: $D_R \leftarrow \varnothing$
2: $confmin \leftarrow 0$
3: **for** each document $d_{Cj} \in D_C$ **do**
4: $supmin \leftarrow compute\_supmin(d_{Cj}, size\_window, \alpha, D, T, W)$
5: $d_{Rj} \leftarrow apriori(d_{Cj}, supmin, confmin)$
6: $D_R \leftarrow D_R \cup d_{Rj}$
7: **end for**
8: Return D_R

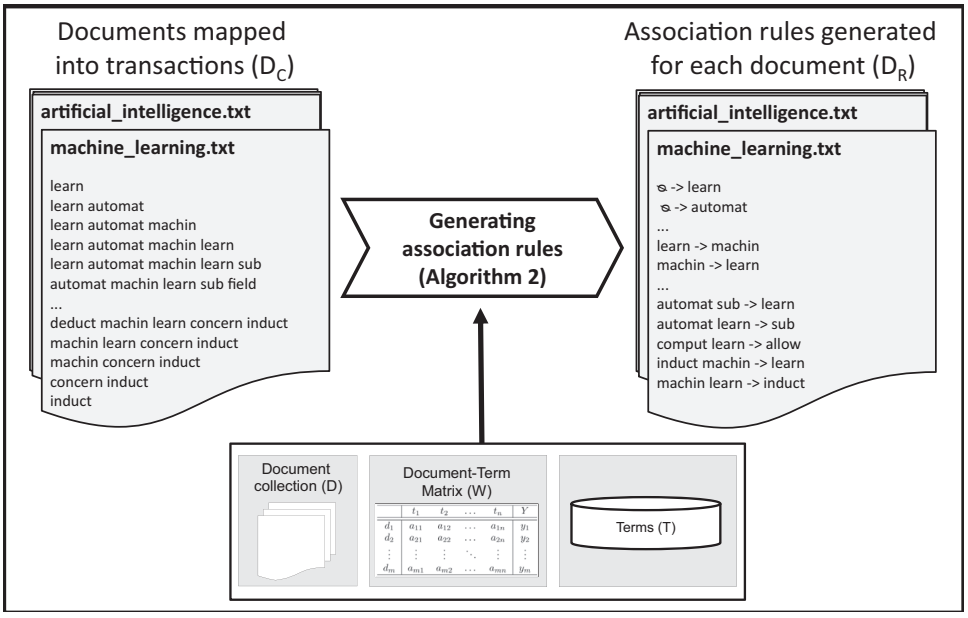

Fig. 3. Obtaining the association rules for each document by using its set of transactions.

Table 3
Objective measures to select association rules.

Measure	Description
Support	It is an indication of how frequently the itemset appears in the dataset.
Confidence	It is an indication of how often the rule has been found to be true.
ϕ-Coefficient	Indicates the degree of correlation between two variables (e.g., antecedent and consequent of a rule).
Novelty	Measures the percentage of additional transactions covered by an association rule that is higher than expected.
Gini Index	Measures the decrease in impurity or uncertainty of a target variable (e.g., consequent), conditioned to the knowledge of the value of a predictive variable (e.g., antecedent). The higher the value of the measure, the stronger will the association between the variables be.
Certainty Factor	Measures how much the belief in the consequent increases with the observation of the antecedent.
Laplace	It is a variation of the Confidence measure that penalizes very specific rules (i.e., rules that cover few transactions).
Kappa	It is a coefficient of agreement.
Added Value	Indicates how frequently the consequent increases in the presence of the antecedent.
Klosgen	Represents the combination of the measures Added Value and Support.
J-Measure	Represents the product of two values. In our case, the probability $P(LHS)$ and $P(RHS\|LHS)$.
Collective Strength	Measures the correlation between a set of items.
Mutual Information LHS	Measures the amount of reduction in uncertainty of the consequent when the antecedent is known.
Lift	Measures how far from independency the antecedent and the consequent are.
IS	Represents the geometric mean between Lift and Support measures.
Lambda	Measures the relative decrease in the probability of an error when calculating an attribute while taking into account the presence or absence of another attribute.
Conviction	It is somewhat inspired in the logical definition of implication and attempts to measure the degree of implication of a rule.

In order to reduce user effort, we have adopted a heuristic method to calculate the cut-off value for a given objective measure in each document. The heuristic involves calculating the average value of the objective measure while considering all rules generated for the document (line 5), using the formula:

$$min\_measure(d_{Rj}) = \frac{\sum_{\forall r_{jq} \in d_{Rj}} value\_measure(r_{jq})}{|d_{Rj}|}, \quad (2)$$

where r_{jq} is the qth association rule in d_{Rj}, $value\_measure(r_{jq})$ is the value of the objective measure for the association rule r_{jq} and $|d_{Rj}|$ is the number of association rules in d_{Rj}. This heuristic has already been used successfully by Rossi and Rezende (2011). With the defined cut-off value, all association rules with an objective measure value greater than the cut-off value are selected by the algorithm (lines 6 to 10). The result is the collection $D_S = \{d_{S1}, d_{S2}, \ldots, d_{Sm}\}$, where d_{Sj} corresponds to the association rules selected from d_{Rj}, in which the objective measure value is greater than *min_measure*.

3.2. Identifying general context of the relation

Once we have obtained terms considering their relation in the local context, we cluster them in order to have, in a same group, the terms that are related to the same subject or topic. To create these groups, it is necessary to identify a relation among terms at the collection level, so that we can discover the topics from the collection. However, information about existing topics in the collection is unknown. Thus, we consider that there is a set of terms more likely to be used to express a particular topic or subject in each document, and that each term is used in a document to

Algorithm 3 Selectin rules.

Require: Collection of association rules D_R; Objective measure (*objective_measure*) for selecting association rules.
Ensure: The collection of selected association rules D_S.

```
D_S ← ∅
for each document d_Rj ∈ D_R do
    d_Rj ← calculate_value_objective_measure(d_Rj, objective_measure)
    d_Sj ← ∅
    min_measure ← calculate_min_measure(d_Rj, objective_measure)
    for each association rule r_jq ∈ d_Rj do
        if value of objective measure of r_jq > min_measure then
            d_Sj ← d_Sj ∪ r_jq
        end if
    end for
    D_S ← D_S ∪ d_Sj
end for
Return D_S
```

express a particular topic. In this work, we also consider that each document can handle one or more subjects or topics.

In order to discover the unknown information concerning the topics, we propose clustering the association rules obtained from the previous step. Traditional association rule clustering usually employs, as a similarity measure, some variation of the schema which counts the common items between two association rules, either the rule itself or the transactions that the rule covers. However, in the case of association rules obtained from textual documents, there is context information that can be better explored. In our work, we investigate the idea that the transactions covered by each association rule obtained from textual documents correspond to the neighborhood of each term, and that terms with similar neighborhoods are possibly related to the same topic. According to Turney and Pantel (2010), terms that occur in similar contexts tend to have similar meanings. Thus, let **k** be the desired number of topics in a collection. The association rules are clustered in **k** clusters so that each cluster corresponds to a topic, and the association rules of that cluster contains the terms that are most likely to be used when a document describes a particular topic.

To obtain the clusters of association rules, we need a structured representation that can capture the information concerning the neighborhood of each term. To do that, we propose an intermediate representation for the selected association rules, in which the extracted terms of the documents $T = \{t_1, t_2, \ldots, t_n\}$ are used to create an association rule-term matrix W_R. According to our proposal, the weight of the terms for each association rule is obtained by making use of the set of transactions covered by the rule, i.e., the set of transactions that generated the rule. Thus, we consider, in this process, the local context of the relation and its neighborhood in order to discover the general context of the relations. Each rule r_{jq} is represented by a term vector $\vec{r_{jq}} = \{w_{1jq}, w_{2jq}, \ldots, w_{njq}\}$, where each weight w_{ijq} quantifies the importance of the term $t_i \in T$ for the association rule r_{jq}, so that w_{ijq} represents the frequency of the term t_i in the transactions in $d_{Cj} \in D_C$ covered by the rule r_{jq}. Then, these association rules are represented in the association rule-term matrix W_R, in which each row corresponds to the vector of the association rule $\vec{r_{jq}}$, and each column describes the importance w_{ijq} of the term t_i in r_{jq}. Identical association rules obtained from different documents are treated as different rules, since each of them may be related to different topics. Thus, the proposed model presents an explicit mechanism to identify polysemy.

In Fig. 4, we illustrate the construction of the association rule-term matrix W_R. For each association rule, the transactions covered by the rule are found, and each term is counted to indicate its frequency in the rule. This action not only allows us to identify association rules with common terms, but also to represent - for each term - its presence in the neighborhood of the rule.

Once the representation W_R is obtained, a traditional clustering algorithm is applied, as shown in Fig. 1. For this step, we have used the Bi-Secting K-means algorithm, which has provided the best results. As in other methods, the number k of generated clusters must be informed by the user. The result of this process is the set of clusters $Z = \{z_1, z_2, \ldots, z_k\}$, where each cluster $z_x \in Z$ contains association rules obtained from the different documents of D_S.

3.3. Generating document-topic representation

Finally, we calculate the weight of the topics for each document to obtain a final representation format, i.e., the document-topic matrix. In the matrix, each set of rules will represent a topic. In the LARCM, the topic weight for each document is calculated with basis on the proportion of the association rules of the document which are in each cluster. Therefore, the weight of a topic can be seen as the proportion of the document that is represented by the topic. The topic weight for each document is calculated with the following equation:

$$topic\_weight(d_j, z_x) = \frac{\left|d_{Sj} \cap z_x\right|}{\left|d_{Sj}\right|}, \tag{3}$$

where $|d_{Sj} \cap z_x|$ is the total number of association rules of d_{Sj} in the topic z_x and $|d_{Sj}|$ is the total number of association rules in d_{Sj}.

In order to analyze the terms and identify the topic descriptor, a list with the most relevant terms of the topic must be presented to the user. In this work, we propose that such a list be based on the selection of the L best association rules of each cluster $z_x \in Z$, according to an objective measure for association rules. The objective measure will be the same one chosen to identify the local context (Section 3.1). Santos, Rezende, and de Carvalho (2014) concluded that this process is able to select more meaningful descriptors in comparison to those obtained by the LDA model, which is considered the state-of-the-art method in the field. The descriptor selection process can be defined as follows:

$$descriptors(z_x) = \{best_l(z_x), l = 1, \ldots, L\}, \tag{4}$$

where $best_l$ is the function that selects the L best association rules according to the objective measure chosen for the xth topic. L is a fixed number, empirically defined as $L = 10$ in many works in the literature. The result can be presented in a Term-Topic Matrix that represents–for each topic–the list with its relevant terms, as can be seen in Fig. 1.

4. Empirical evaluation

After extracting the topics, we need to evaluate them. The evaluation of topic models is a major challenge, since the process to select the relevant topics in a collection is highly subjective, and depends on the interpretation of many evaluators in order to become meaningful. In this section, we evaluate our proposal based on the automatic evaluation process presented by Lau, Newman, and Baldwin (2014), which consists of a process that simulates an evaluation by experts and provides a fairly reliable quality estimation for the topics.

According to Newman, Lau, Grieser, and Baldwin (2010), the interpretability of a topic must be based on the coherence observed in the n terms selected for the topic descriptors. Following a set of standardized instructions, the extracted topics are sent to a group of experts, who must evaluate the quality of the topics by using a 3-point scale to represent the coherence of the terms. The value 1 means a topic of little usefulness, whereas the value 3 means

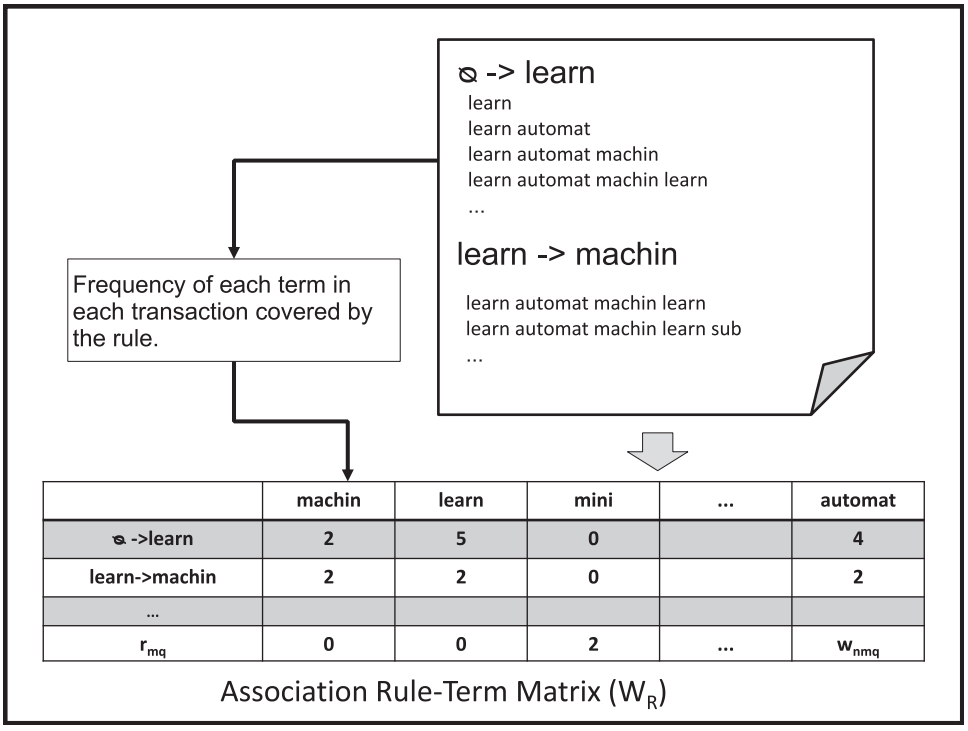

Fig. 4. Process to create the association rule-term matrix W_R.

a useful topic. In Lau et al. (2014), the authors propose the automatic calculation of the coherence measure normalized pointwise mutual information (NPMI). This measure is indicated as the most adequate for the evaluation, since its result is close to the one provided by experts, and can be used to automate the coherence evaluation of the topics, considering the terms selected as topic descriptors and their co-occurrence with respect to a reference collection. In this work, we used the Wikipedia[1] in English as the reference collection to calculate the co-occurrence among the terms selected as descriptors. All articles of the reference collection must be preprocessed following the same procedure used for the documents in the evaluated collection. In Lau et al. (2014), the observed coherence (OC) for the *k*th cluster or topic, considering its set of descriptors, is given by the sum of the NPMI value of all combinations of pairs of terms from the set of descriptors. Thus, the formula for $OC\_NPMI(C_k)$ measure can be defined as:

$$OC\_NPMI(C_k) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} \frac{log(\frac{P(t_j,t_i)}{P(t_i)\cdot P(t_j)})}{-log(P(t_i,t_j))}, \quad (5)$$

where C_k is a set with n terms selected for the topic descriptor; $P()$ and $P(,)$ are functions to calculate the probability and the conditional probability, respectively; $t_i \in C_k$ and $t_j \in C_k$ are two terms selected for the topic descriptor.

The NPMI was proposed with the aim to reduce the bias of the PMI (pointwise mutual information) measure for infrequent terms, and its value is in the range $[-1, 1]$. Thus, we have an easier comparison among the results obtained with the different topic models. With the NPMI, the higher the value of the measure, the better is the consistency of the topic.

4.1. Datasets

To evaluate our proposal, we used the text collections presented in Table 4. For each collection, we present the number of classes (**# classes**), the number of documents (**# docs**), the number of distinct terms (**# distinct terms**), the average number of terms per document (**Avg terms per doc**), and the description of the collection. The collections were preprocessed as follows. A standardization of the texts was carried out by removing numbers, symbols and also the stopwords. The letters with special characters, as for example "ç", "í" and "ã", were replaced by the respective letters without the special characters, i.e., "c", "i" and "a". The terms were obtained by applying the algorithm of Porter (Nogueira et al., 2008; Porter, 1997) to reduce the words to their stem. Since the topic modeling approaches reduce dimensionality, we did not utilize any attribute selection process.

4.2. Experimental setup

In this section we present the setup for the experimental evaluation of the LARCM approach. As a baseline, we used the traditional LDA, a well-known technique for topic extraction models. The topics of the LDA model were produced from a bag-of-words formed by simple terms and also from the representation obtained by the bag-of-related-words consisting of simple and correlated terms (Rossi & Rezende, 2011). The use of the bag-of-related-words as input by the LDA model allows us to compare the impact of using correlated terms in relation to the traditional LDA model. The extraction of correlated terms prior to the execution of the LDA is highlighted by Lau et al. (2013) as a viable alternative for including term dependency in the process, in order to obtain better results.

The bag-of-related-words was generated by using the FEATuRE tool[2]. Based on the results presented by Rossi and Rezende (2011), we decided to use the automatic support and the average value of the objective measure to select the attributes (i.e., the terms). We also used the objective measure Kappa to construct the bag-of-related-words representation for this evaluation. For the construction of the transactions, the size-5 window was adopted, which

[1] https://en.wikipedia.org/wiki/Main_Page.

[2] http://sites.labic.icmc.usp.br/feature.

Table 4
Description of the text collections.

Collection	# classes	# docs	# distinct terms	Avg terms per doc	Description
ACM-1	5	399	40,918	4337	Collection of scientific articles from conferences of different computing areas extracted from ACM digital repository (Rossi & Rezende, 2011)
ACM-2	5	410	47,907	3516	
ACM-3	5	416	40,181	4389	
ACM-4	5	394	53,474	4499	
ACM-5	5	471	40,732	4487	
ACM-6	5	437	54,088	5033	
ACM-7	5	469	55,015	5980	
ACM-8	5	495	50,486	5807	
Re8	8	7674	17,335	58	Collection of news articles from the 8 more frequent classes of Reuters-21578[a]

[a] http://www.daviddlewis.com/resources/testcollections/reuters21578.

Table 5
Parameter values used by the topic modeling approaches.

LDA	
Input data representation	bag-of-words bag-of-related-words (objective measure: *kappa* / window size: 5)
Hyperparameters α and β	estimated automatically by the tool
Iterations for the Gibbs sampling	1000
Number of topics (k)	50, 100, 150
LARCM	
Window size	5
Minimum support / confidence per document	supmin defined by Eq. (1) / confmin = 0
Objective measures	Support, Confidence, ϕ-Coefficient, Novelty, Gini Index, Certainty Factor, Laplace, Kappa, Added Value, Klosgen, J-Measure, Collective Strength, Mutual Information LHS, Lift, IS, Lambda, Conviction
Cut-off value for objective measure per document	defined by Eq. (2)
Clustering algorithm	Bi-Secting K-means with the cosine similarity measure
Number of clusters (k)	50, 100, 150

presented the best results according to Rossi and Rezende (2011). The LDA models were generated by using the MALLET tool[3].

According to Liu (2011), only the rules with 2 or 3 terms are required, since correlated terms composed by more than 3 terms are infrequent. Thus, we choose to generate association rules for the LARCM with a maximum of 3 terms. The transactions are generated by using a size-5 window, the same value used to obtain the bag-of-related-words representation for the LDA. Due to the association rule algorithm[4] used in this evaluation, the association rules will have some items in the antecedent, but only one item in the consequent.

The association rules are generated for each document in the collection, and a subset of these rules are selected by using one of the objective measures presented in Table 5, according to the process described in Section 3.1. For clustering the rules, we use the algorithm Bi-Secting K-means[5] with the cosine similarity measure. For each cluster, the ten best rules are selected, according to the objective measure, to create the set of descriptors.

We compare the LARCM against the LDA for each objective measure presented in Table 5. Following Chang et al. (2009), we evaluated the topic descriptors by using the values of $k = 50$, $k = 100$ and $k = 150$ for both LDA and LARCM models. For each topic, we evaluate the ten terms most likely to generate the set of descriptors. To run the process proposed by Lau et al. (2014), we used the tool topic-interpretability[6]. The reference collection used was Wikipedia in English, version extracted on January 15th, 2015.

[3] http://mallet.cs.umass.edu.
[4] http://www.borgelt.net/apriori.html.
[5] http://glaros.dtc.umn.edu/gkhome/views/cluto.
[6] https://github.com/jhlau/topic_interpretability.

4.3. Results

After calculating the observed coherence measure (i.e., $OC\_NPMI(C_k)$) for each topic and model, we select the sets with the 25% best topics. For example, for a model generated with a value of k equal to 100 topics, the set will consist of the 25 topics with the best observed coherence value. From this set, the topic with the highest (Best OC) and the lowest (OC in 25%) values are considered for evaluation.

In this evaluation, our goal is to verify if the LARCM produces topics with better interpretability than those produced by the LDA model. Topics with better interpretability are those that facilitate user interaction with the document collection in an exploratory way.

The values of observed coherence for the best topic and the lowest value of the measure for the topic in the upper quartile are presented in Tables 6 and 7. We calculated the LARCM values for all objective measures presented in Table 5.

By looking at the tables, we can see that it was not possible to indicate one measure as the most interesting for all cases. By analyzing only the Best OC values (Table 6), the LARCM obtained topics with better OC values compared to LDA models in 17 out of 27 evaluated cases (i.e., 3 values of k and 9 collections of texts). In the 10 cases where the LDA leads to its best OC values, the LARCM presents similar OC values. On the other hand, when the LARCM is superior to LDA, the values obtained are significantly higher. In the ACM-2 collection with $k = 150$, our proposal–combined with the Lift measure–yields the most significant difference in relation to the best topic produced by the LDA. Additionally, for the observed coherence value in 25% (OC in 25%) in Table 7, the LDA model does slightly better than the LARCM. In general, the objective measures

Table 6
Observed Coherence (OC) value for the topic with the best evaluation (Best OC). The highest value for each collection is in boldface.

Combination	ACM-1	ACM-2	ACM-3	ACM-4	ACM-5	ACM-6	ACM-7	ACM-8	Re8
***k*= 50**									
LDA + *bag-of-words*	**0.35**	0.34	**0.34**	0.33	0.34	0.36	0.33	**0.33**	0.28
LDA + *bag-of-related-words*	0.32	0.34	0.30	0.31	0.33	0.34	**0.34**	0.31	**0.30**
LARCM + Support	0.14	0.14	0.12	0.20	0.26	0.19	0.18	0.19	0.29
LARCM + Confidence	0.20	0.27	0.15	0.18	0.19	**0.38**	0.24	0.18	0.23
LARCM + ϕ-Coefficient	0.18	0.17	0.27	0.20	0.28	0.18	0.26	0.20	0.21
LARCM + Novelty	0.21	0.22	0.18	0.24	0.33	0.33	0.22	0.23	0.29
LARCM + Gini Index	0.20	0.26	0.18	0.22	**0.39**	0.33	0.27	0.26	0.25
LARCM + Certainty Factor	0.20	0.16	0.16	0.19	0.21	0.28	0.24	0.17	0.25
LARCM + Laplace	0.16	0.27	0.21	0.16	0.16	0.33	0.24	0.18	0.24
LARCM + Kappa	0.16	**0.38**	0.26	0.36	0.23	0.33	0.24	0.20	0.23
LARCM + Added Value	0.15	0.23	0.16	0.21	0.21	0.20	0.28	0.17	0.25
LARCM + Klosgen	0.15	0.29	0.16	0.27	0.26	0.25	0.26	0.23	0.22
LARCM + J-Measure	0.25	0.29	0.29	0.25	0.34	0.27	0.26	0.29	0.25
LARCM + Collective Strength	0.22	0.28	0.23	0.33	0.27	0.19	0.27	0.28	0.29
LARCM + Mutual Information LHS	0.26	0.27	**0.34**	0.25	0.25	0.26	0.26	0.26	0.25
LARCM + Lift	0.20	0.28	0.31	0.25	0.21	0.24	0.20	0.21	0.23
LARCM + IS	0.27	0.26	0.25	0.29	0.25	0.33	0.29	0.27	0.25
LARCM + Lambda	0.11	0.12	0.12	0.14	0.12	0.10	0.08	0.18	0.20
LARCM + Conviction	0.21	0.20	0.22	**0.38**	0.27	0.29	0.18	0.24	0.23
***k*= 100**									
LDA + *bag-of-words*	**0.36**	0.36	0.34	0.33	0.34	0.33	0.33	**0.34**	0.30
LDA + *bag-of-related-words*	0.35	0.34	0.33	0.31	0.35	0.32	0.31	0.33	**0.31**
LARCM + Support	0.21	0.23	0.29	0.21	0.26	0.18	**0.41**	0.21	0.25
LARCM + Confidence	0.18	0.27	0.18	0.23	0.24	0.33	0.24	0.18	0.24
LARCM + ϕ-Coefficient	0.23	0.26	0.22	0.29	0.26	0.18	0.26	0.28	0.23
LARCM + Novelty	0.23	0.32	0.26	0.27	0.29	0.30	0.30	0.24	0.23
LARCM + Gini Index	0.25	0.31	0.25	0.26	0.26	0.33	0.22	0.24	0.25
LARCM + Certainty Factor	0.22	0.43	0.22	0.29	0.33	**0.40**	0.22	0.20	0.26
LARCM + Laplace	0.25	0.26	0.20	0.24	0.32	0.38	0.24	0.18	0.26
LARCM + Kappa	0.31	0.28	0.36	0.36	**0.43**	0.33	0.25	0.22	0.25
LARCM + Added Value	0.24	**0.48**	0.21	0.21	0.41	0.36	0.24	0.19	0.28
LARCM + Klosgen	0.21	0.34	0.36	0.30	0.28	0.33	0.20	0.23	0.23
LARCM + J-Measure	0.29	0.26	0.30	0.27	0.28	0.30	0.26	0.27	0.24
LARCM + Collective Strength	0.30	0.25	0.25	0.33	0.39	0.25	0.28	0.28	0.28
LARCM + Mutual Information LHS	0.27	0.32	0.29	0.26	0.26	0.29	0.27	0.31	0.24
LARCM + Lift	0.33	0.44	0.25	0.28	0.27	0.36	0.21	0.26	0.25
LARCM + IS	0.25	0.25	**0.37**	0.26	0.29	0.34	0.24	0.25	0.25
LARCM + Lambda	0.18	0.14	0.19	0.21	0.21	0.23	0.12	0.17	0.20
LARCM + Conviction	0.25	0.26	0.34	**0.37**	0.29	0.35	0.36	0.27	0.24
***k*= 150**									
LDA + *bag-of-words*	**0.36**	0.34	0.35	0.34	0.35	0.35	0.33	**0.32**	0.32
LDA + *bag-of-related-words*	0.34	0.35	0.35	0.32	0.36	0.35	0.31	**0.32**	0.31
LARCM + Support	0.23	0.25	0.32	0.29	0.32	0.21	0.31	0.28	0.27
LARCM + Confidence	0.27	0.35	0.25	0.23	0.33	0.38	0.24	0.18	0.27
LARCM + ϕ-Coefficient	0.32	0.26	0.27	0.32	**0.40**	0.33	0.33	0.31	0.34
LARCM + Novelty	0.29	0.39	0.33	0.27	0.38	0.33	**0.36**	0.23	0.34
LARCM + Gini Index	0.22	0.43	0.24	0.25	0.38	**0.42**	0.25	0.25	0.29
LARCM + Certainty Factor	0.28	0.31	0.35	0.27	0.24	0.35	0.27	0.21	0.27
LARCM + Laplace	0.23	0.35	0.26	0.23	0.32	0.33	0.24	0.20	0.27
LARCM + Kappa	0.23	0.38	0.31	0.36	0.34	0.39	0.25	0.24	0.28
LARCM + Added Value	0.23	0.42	0.27	0.24	0.41	0.33	0.28	0.24	0.35
LARCM + Klosgen	0.21	0.42	0.24	0.25	0.39	0.36	0.35	0.21	0.34
LARCM + J-Measure	0.24	0.32	0.27	0.26	0.30	0.33	0.26	0.28	0.29
LARCM + Collective Strength	0.27	0.35	0.28	0.33	0.33	**0.42**	0.23	0.28	**0.44**
LARCM + Mutual Information LHS	0.26	0.33	**0.40**	0.30	0.29	0.37	0.27	0.25	0.31
LARCM + Lift	0.32	**0.48**	0.27	0.24	0.25	0.35	0.22	0.28	0.30
LARCM + IS	0.30	0.26	0.26	0.30	0.27	**0.42**	0.33	0.27	0.32
LARCM + Lambda	0.19	0.21	0.18	**0.46**	0.21	0.27	0.19	0.21	0.31
LARCM + Conviction	0.33	0.25	0.24	0.32	0.32	0.28	**0.36**	0.24	0.29

that provide a bigger coverage for the association rules seem to be more appropriate for selecting the descriptors.

Considering only the best OC values obtained with the LDA model in Tables 6 and 7, the representation that uses the bag-of-words as input was superior to the model that uses the bag-of-related-words. On the other hand, considering the value of observed coherence in 25% (OC in 25%), the model using bag-of-related-words presented similar results to those of the traditional LDA–which does not justify its usage, given the high cost of obtaining this representation if compared to the bag-of-words representation.

Finally, we present some examples of terms from topic descriptors in Table 8. There, we can observe meaningful correlated terms selected as topic descriptors by the LARCM model, such as *sensorized wireless, wireless network, federal budget* and *deficit reduction*. These correlated terms make the topic descriptors more interpretable than the ones built with only single terms by the LDA models. Besides, single terms, like *network* and *federal*, can be frequently related to other single terms in an inappropriated way, decreasing the interpretability of a topic descriptor. This fact emphasize the importance of the correlated terms generated by the LARCM model.

Table 7
Observed Coherence (OC) value for the topic with the lowest value among the 25% best topics (OC in 25%). The highest value for each collection is in boldface.

Combination	ACM-1	ACM-2	ACM-3	ACM-4	ACM-5	ACM-6	ACM-7	ACM-8	Re8
***k*= 50**									
LDA + *bag-of-words*	**0.27**	**0.28**	**0.27**	**0.26**	0.27	**0.28**	**0.27**	**0.27**	0.22
LDA + *bag-of-related-words*	0.26	**0.28**	**0.27**	**0.26**	**0.28**	0.27	**0.27**	0.26	**0.24**
LARCM + Support	0.06	0.06	0.07	0.05	0.07	0.05	0.07	0.06	0.09
LARCM + Confidence	0.04	0.09	0.06	0.06	0.08	0.05	0.06	0.05	0.09
LARCM + ϕ-Coefficient	0.06	0.10	0.09	0.05	0.08	0.07	0.07	0.06	0.10
LARCM + Novelty	0.08	0.09	0.08	0.09	0.08	0.08	0.08	0.08	0.15
LARCM + Gini Index	0.08	0.11	0.09	0.09	0.09	0.08	0.08	0.06	0.16
LARCM + Certainty Factor	0.05	0.07	0.06	0.06	0.07	0.05	0.06	0.05	0.16
LARCM + Laplace	0.04	0.08	0.06	0.06	0.07	0.05	0.06	0.05	0.10
LARCM + Kappa	0.09	0.10	0.05	0.08	0.08	0.08	0.06	0.07	0.16
LARCM + Added Value	0.07	0.08	0.05	0.07	0.09	0.04	0.05	0.04	0.16
LARCM + Klosgen	0.05	0.09	0.07	0.07	0.07	0.06	0.06	0.06	0.14
LARCM + J-Measure	0.20	0.22	0.19	0.20	0.20	0.21	0.20	0.19	0.18
LARCM + Collective Strength	0.08	0.12	0.09	0.10	0.08	0.07	0.06	0.07	0.15
LARCM + Mutual Information LHS	0.21	0.21	0.20	0.20	0.20	0.21	0.21	0.20	0.18
LARCM + Lift	0.11	0.10	0.08	0.12	0.08	0.09	0.07	0.07	0.16
LARCM + IS	0.19	0.21	0.19	0.17	0.18	0.21	0.19	0.18	0.17
LARCM + Lambda	0.05	0.06	0.05	0.04	0.04	0.04	0.03	0.03	0.12
LARCM + Conviction	0.07	0.11	0.13	0.12	0.08	0.09	0.09	0.09	0.16
***k*= 100**									
LDA + *bag-of-words*	**0.26**	**0.27**	**0.26**	**0.25**	**0.27**	**0.26**	**0.26**	**0.26**	**0.25**
LDA + *bag-of-related-words*	**0.26**	0.26	**0.26**	**0.25**	0.26	**0.26**	**0.26**	0.25	0.23
LARCM + Support	0.06	0.07	0.07	0.06	0.06	0.06	0.06	0.07	0.13
LARCM + Confidence	0.07	0.09	0.07	0.07	0.08	0.07	0.07	0.05	0.11
LARCM + ϕ-Coefficient	0.06	0.10	0.09	0.08	0.09	0.08	0.07	0.06	0.10
LARCM + Novelty	0.09	0.13	0.11	0.10	0.12	0.08	0.08	0.06	0.16
LARCM + Gini Index	0.11	0.14	0.09	0.09	0.13	0.10	0.08	0.09	0.16
LARCM + Certainty Factor	0.09	0.12	0.09	0.10	0.10	0.09	0.09	0.07	0.17
LARCM + Laplace	0.07	0.09	0.07	0.08	0.07	0.06	0.08	0.06	0.11
LARCM + Kappa	0.09	0.12	0.11	0.10	0.11	0.09	0.09	0.08	0.16
LARCM + Added Value	0.09	0.11	0.10	0.09	0.09	0.08	0.08	0.06	0.16
LARCM + Klosgen	0.09	0.12	0.10	0.08	0.09	0.09	0.07	0.07	0.17
LARCM + J-Measure	0.19	0.19	0.18	0.17	0.18	0.18	0.18	0.17	0.18
LARCM + Collective Strength	0.11	0.14	0.10	0.11	0.09	0.11	0.09	0.07	0.15
LARCM + Mutual Information LHS	0.19	0.19	0.19	0.19	0.20	0.19	0.20	0.19	0.18
LARCM + Lift	0.12	0.15	0.12	0.13	0.13	0.12	0.11	0.10	0.17
LARCM + IS	0.15	0.18	0.17	0.15	0.17	0.18	0.18	0.16	0.18
LARCM + Lambda	0.05	0.07	0.06	0.05	0.05	0.05	0.05	0.04	0.13
LARCM + Conviction	0.11	0.13	0.14	0.13	0.11	0.13	0.10	0.10	0.16
***k*= 150**									
LDA + *bag-of-words*	**0.26**	**0.26**	0.25	**0.25**	**0.26**	**0.26**	**0.25**	**0.25**	**0.24**
LDA + *bag-of-related-words*	0.25	**0.26**	**0.26**	**0.25**	0.25	0.25	**0.25**	**0.25**	0.22
LARCM + Support	0.06	0.07	0.06	0.07	0.06	0.06	0.06	0.06	0.13
LARCM + Confidence	0.07	0.10	0.08	0.07	0.08	0.06	0.07	0.06	0.10
LARCM + ϕ-Coefficient	0.06	0.09	0.07	0.07	0.08	0.08	0.07	0.06	0.12
LARCM + Novelty	0.10	0.13	0.11	0.11	0.11	0.11	0.10	0.08	0.16
LARCM + Gini Index	0.11	0.13	0.10	0.11	0.12	0.11	0.11	0.09	0.16
LARCM + Certainty Factor	0.11	0.12	0.10	0.11	0.10	0.10	0.08	0.07	0.16
LARCM + Laplace	0.08	0.09	0.08	0.08	0.08	0.06	0.07	0.06	0.10
LARCM + Kappa	0.12	0.13	0.11	0.11	0.10	0.10	0.10	0.08	0.16
LARCM + Added Value	0.10	0.11	0.10	0.11	0.10	0.10	0.10	0.07	0.16
LARCM + Klosgen	0.09	0.11	0.11	0.10	0.10	0.09	0.09	0.08	0.16
LARCM + J-Measure	0.17	0.18	0.16	0.15	0.17	0.17	0.17	0.16	0.17
LARCM + Collective Strength	0.11	0.12	0.11	0.11	0.11	0.11	0.10	0.09	0.16
LARCM + Mutual Information LHS	0.17	0.18	0.17	0.18	0.17	0.18	0.17	0.18	0.17
LARCM + Lift	0.13	0.14	0.13	0.13	0.13	0.11	0.11	0.11	0.17
LARCM + IS	0.15	0.16	0.15	0.13	0.15	0.15	0.15	0.14	0.17
LARCM + Lambda	0.10	0.07	0.06	0.07	0.06	0.05	0.05	0.06	0.13
LARCM + Conviction	0.06	0.12	0.12	0.12	0.12	0.12	0.10	0.10	0.17

5. LARCM in classification and recommendation applications

The topics obtained by the LARCM can be used for several purposes. In this section, we use them to improve text classification as well as page recommendation applications. Concerning the former, we use the topics to improve document representation in the text classification task. The latter application consists of using the topics as contextual information in order to improve page recommendation by context-aware recommender systems.

5.1. Applying the LARCM in text classification

The text classification task aims to automatically identify the main topics in a document in order to link this document to one or more predefined categories (da Silva Conrado, Gutiérrez, & Rezende, 2012). Text classification is a very important task in the text mining field of research, because it helps to keep large textual document collections organized, thus making it possible to carry out other automatic and manual tasks more quickly.

Table 8
Examples of topic descriptors with at most 10 terms.

Collection	Combination	OC value	Descriptor
ACM-1	LDA + *bag-of-words* (k=150)	0.36	atoms, molecul, information, amino, visualization, acid, structure, sequenc, molecular
ACM-1	LDA + *bag-of-related-words* (k=150)	0.34	node, handoff, address, mobility, message, router, latency, packet, network, mobility node
ACM-1	LARCM + Conviction (k=150)	0.33	wireless voice, sensorized wireless, wireless, wireless network, control congestion, wireless distributed, wireless lan, antenna multi, network wireless
Re8	LDA + *bag-of-words* (k=150)	0.32	chairman, president, vice, officer, company, executive, directors, reuter, board, chief
Re8	LDA + *bag-of-related-words* (k=150)	0.31	budget, domestic, deficit, economic, reagan, economy, policy, administration, national, growth
Re8	LARCM + Collective Strength (k=150)	0.44	cut deficit, federal budget, budget federal, deficit cut, deficit, reduction deficit, deficit budget, volcker, deficit reduction, budget deficit

In order for a better document classification to be achieved, it is necessary to work with terms that conceptually represent the textual document collection. According to Liu (2011), topic models can be used to represent document collections in the text classification task. Thus, in this section we evaluate how the LARCM can contribute to represent document collections in order to improve the accuracy of document classification task.

To provide a fair evaluation, we used traditional classification algorithms that are based on different paradigms (Aggarwal & Zhai, 2012; Sebastiani, 2002). The algorithms are: i) Naive Bayes and Multinomial Naive Bayes (probabilistic paradigm), ii) J48[7] (symbolic paradigm), iii) SMO[8] (statistical paradigm), and iv) IBk[9] (instance-based paradigm).

5.1.1. Datasets

In order to evaluate our proposal in the text classification task, we used the nine text collections presented in Section 4.1. All text collections were pre-processed using the same method described in that section.

5.1.2. Experimental setup

To run the evaluation, we used the Weka tool[10] - version 3.7.10 (Hall et al., 2009). For most algorithms, we adopted the standard parameters that are suggested by Weka. The algorithms SMO and IBk are very sensitive to the initialization of their parameters (Batista & Silva, 2009; Braga, 2014), so a few variations were introduced before these algorithms were executed. For the SMO algorithm, experiments were performed using the generalization parameter "c" with its value equal to 1 (Weka standard value) and 10 (chosen empirically based on the analysis of the classification results obtained by Rossi, Marcacini, and Rezende (2013) in various document collections). For the algorithm IBk, following the statements of Batista and Silva (2009), the values of ℸ neighbors were defined in 3, 5 and 7. All algorithms were executed and evaluated taking into account their accuracy, that can be measured as follows:

$$Acur = |D_{cc}|/|D|, \tag{6}$$

where *Acur* is the proportion of documents correctly classified $|D_{cc}|$ in relation to the total number of documents in the collection $|D|$.

To calculate the accuracy, we use the 10-fold cross-validation protocol, where each classifier is trained 10 times. Conventionally, the average accuracy obtained from the 10 runs is used to represent the accuracy of the classifier for a representation (set of topics) obtained with the LARCM. We calculate the mean measure and its standard deviation for each representation, and then we apply the *t-Student* test to statistically evaluate the difference among the means.

However, we have several representation models (*mr*) applied to different classifiers (*cl*), in different document collections (*cd*) and with different numbers of topics (*k*). In this case, the analysis of variance must be able to evaluate the differences among the several LARCMs, considering the effects of the other components. Thus, for a fair evaluation, the total variance must be decomposed considering all these factors (Moura, 2009). This decomposition of variance can be represented in the following linear model (for details see Searle, 1971):

$$\widehat{m(Acur_{mr})} = \hat{\mu} + \hat{cd} + \hat{k} + \hat{mr} + \hat{cl} + \hat{e}, \tag{7}$$

where:

- $\widehat{m(Acur_{mr})}$: estimated accuracy value for the representation model *rm* (LDA model with bag-of-words or bag-of-related-words, and LARCMs with one of the objective measures);
- $\hat{\mu}$: estimated value for the general mean of the accuracy;
- $\hat{cd}$: estimated value for the influence of the text collection *cd* on the estimated accuracy value;
- $\hat{k}$: estimated value for the influence of the number of topics *k* on the estimated accuracy value;
- $\hat{mr}$: estimated value for the influence of the representation model *mr* on the estimated accuracy value;
- $\hat{cl}$: estimated value for the influence of the classifier in the estimated accuracy value;
- $\hat{e}$: estimated value of the model error in the representation model *rm*.

By using the decomposition, a general mean is obtained for each representation by considering all results in all document collections and for all trained classifiers. Thus, we can compare the influence of each representation model on the improvement of classifier performances in relation to a general average. After considering the influence of different classifiers on different document collections with different numbers of topics, we can compare the noises from these inferences. It is therefore possible to identify the models that will contribute statistically to an improvement in the quality of a classification task.

[7] The implementation of the decision tree algorithm C4.5 in Weka.
[8] The implementation of the SVM algorithm in Weka.
[9] The implementation of KNN in Weka.
[10] http://www.cs.waikato.ac.nz/ml/weka.

Table 9
Comparing the average accuracy of the classifiers for each representation model.

Model	Accuracy	Group
LARCM + Support	84.900	*a*
LDA + *bag-of-words*	83.680	*b*
LARCM + ϕ-Coefficient	81.133	*c*
LARCM + Novelty	80.554	*c d*
LARCM + Gini Index	80.302	*c d e*
LARCM + Certainty Factor	80.263	*c d e*
LARCM + Confidence	80.231	*c d e*
LARCM + Laplace	80.055	*d e*
LARCM + Kappa	79.917	*d e*
LARCM + Added Value	79.840	*d e*
LARCM + Klosgen	79.811	*d e*
LARCM + J-Measure	79.623	*d e*
LARCM + Collective Strength	79.533	*d e*
LARCM + Mutual Information LHS	79.469	*d e*
LARCM + IS	79.457	*d e*
LARCM + Lift	79.209	*e*
LDA + *bag-of-related-words*	78.378	*f*
LARCM + Lambda	77.894	*f*
LARCM + Conviction	68.830	*g*

5.2. Results

As already stated, the goal of this evaluation is to assess the performance of the LARCM in representing document collections, in order to improve the accuracy of the document classification task. Since this task is of the highest importance in the text mining literature, it is imperative that any proposed representation be capable of improving the results obtained by other well-established representations, or at least of approaching these results while offering other indirect advantages.

In this section we summarize all results (the accuracy for each classification algorithm is presented in Appendix A). Table 9 shows the average accuracy for each representation model in the classification task. The accuracy value presented in the table corresponds to the average obtained considering all classification algorithms, the quantity of topics k, and document collections evaluated for each topic model. This average accuracy is calculated by decomposing all the factors involved in the process by using the linear model, as described in the previous section. Thus, the mean value already takes into account any possible bias from the classification algorithm, as well as the impact of the collections on the representations, and the possible effects regarding the number of topics. In the table, the first column identifies the evaluated model and the second one refers to the average accuracy as described above. The third column is obtained by applying the statistical test. Models that are in a same group, indicated by the letters (a, b, c, d,...), do not present statistically significant difference, and therefore, they present similar performance. When the models are related to different letters in the last column, this fact indicates that there is a statistically significant difference among them and, in this case, we can identify the best model for the classification task.

As seen in Table 9, the LARCM, combined with the objective measure Support, obtained the best average accuracy, with statistically significant difference from all other models. Support measures the degree of association between the antecedent and the consequent of an association rule. It thus favors the extraction of terms that are strongly dependent. For most of the other objective measures, we have a similar impact on the classification task, and given that most measures are in a same group (i.e., group "d"), there is no statistically significant difference among them.

For the LDA model, we obtained the best results with the bag-of-words, which yields results with statistically significant difference with respect to almost all other combinations, i.e., it is the only model in group "b" (see Table 9). On the other hand, the LDA model with bag-of-related-words did not perform well at the classification task, when compared to the other models. In our experiment, the addition of correlated terms in the LDA model does not seem to bring so many benefits. This result provides evidence that, for a model to provide significant gains by using correlated terms, it should use the dependency information among terms in an explicit way during the topic identification process.

5.3. Applying the LARCM in context-aware recommender systems

Nowadays, most web sites offer their users a large number of items (e.g., movies, music, web pages, etc.). Finding relevant content that appeals to individual tastes has, therefore, become a challenge. Recommender systems have emerged in response to this problem. A recommender system is an information filtering technology which can be used to output a personalized ranking of items that are likely to be of interest to the user.

A context-aware recommender system can make recommendations by incorporating available contextual information into the recommendation process as explicit additional categories of data (Adomavicius, Sankaranarayanan, Sen, & Tuzhilin, 2005). Contextual information is defined as any information that can be used to characterize the situation of an entity (e.g., a web page) (Dey, 2001) and, according to Hariri, Mobasher, Burke, and Zheng (2011), topics can be used as contextual information for context-aware recommender systems. Thus, in this section, we evaluate the effects of using the contextual information (i.e., topics), obtained with the LARCM, in three different context-aware recommender systems. The recommenders are described as follows:

- **Combined reduction:** In Adomavicius et al. (2005), the combined reduction approach (cReduction) uses the contextual information as label to segment the data. A segment is defined as a subset of the overall data selected according to the context or combination of its values. Then, the contextual segments are used as input by recommender systems;
- **Weight PoF and Filter PoF:** Panniello and Gorgoglione (2012) proposed two contextual approaches. The approaches compute the probability of the user's access items in a given context. Then, the probability is used to reorder (weightPoF) or filter out (filterPoF) the recommendations.

We combined the three contextual recommendation strategies with the item-based collaborative filtering (IBCF) algorithm (Deshpande & Karypis, 2004) in order to generate the contextual recommendations. In this algorithm, the recommender model is a matrix representing the similarities between all the pairs of items according to a similarity measure (in our case, the cosine angle). The top-N recommendations are generated with basis on the 4 most similar items (the 4 nearest neighbors). This value provided the best results for our experiments. For the filterPoF algorithm, we used 0.1 as a threshold to filter out the recommendations, since this value provided the best results.

5.3.1. Datasets

The dataset[11] used in the experiments is from the *Brazilian Embrapa Agency of Technology Information* website[12]. This dataset consists of 4,659 users, 15,037 accesses and 1,543 Web pages about agribusiness, written in Portuguese. The textual content of the pages, crawled from the dataset, was used by the LARCM to obtain the contextual information, i.e., topics. To obtain the topics, we used the same parameters presented in Table 5.

[11] To request a copy of the dataset, please contact the corresponding author.
[12] http://www.agencia.cnptia.embrapa.br.

Table 10
Comparing the context-aware recommenders against the IBCF algorithm. The contextual information (i.e., topics) was obtained by using the LARCM and the LDA model, while considering the value of k equal to 50.

	MAP @ 5				MAP @ 10			
	IBCF	cReduction	weightPoF	filterPoF	IBCF	cReduction	weightPoF	filterPoF
LARCM + ϕ-Coefficient	0.2991	0.3727	0.3933	0.1095	0.3089	0.3875	0.4061	0.1142
LARCM + Added Value	0.2991	0.3826	0.3916	0.1405	0.3089	0.3944	0.4033	0.1437
LARCM + Certainty Factor	0.2991	0.3670	0.3937	0.1415	0.3089	0.3819	0.4049	0.1445
LARCM + Collective Strength	0.2991	0.3845	0.3947	0.1401	0.3089	0.3967	0.4059	0.1448
LARCM + Confidence	0.2991	0.3745	0.3880	0.1382	0.3089	0.3884	0.3992	0.1417
LARCM + Conviction	0.2991	0.3723	0.3789	0.1035	0.3089	0.3876	0.3932	0.1093
LARCM + Gini Index	0.2991	0.3714	0.3837	0.1053	0.3089	0.3852	0.3965	0.1104
LARCM + IS	0.2991	0.3821	0.3881	0.1380	0.3089	0.3958	0.4013	0.1414
LARCM + J-Measure	0.2991	0.3658	0.3762	0.1206	0.3089	0.3797	0.3898	0.1242
LARCM + Kappa	0.2991	**0.3976***	**0.4149***	0.1332	0.3089	**0.4069***	**0.4249***	0.1382
LARCM + Klosgen	0.2991	0.3700	0.3890	0.1249	0.3089	0.3799	0.3995	0.1284
LARCM + Lambda	0.2991	0.3932	0.4019	0.1589	0.3089	0.4059	0.4143	0.1617
LARCM + Laplace	0.2991	0.3827	0.3855	0.1109	0.3089	0.3958	0.3954	0.1138
LARCM + Lift	0.2991	0.3841	0.3941	0.1478	0.3089	0.3967	0.4058	0.1529
LARCM + Mutual Information	0.2991	0.3893	0.3966	0.1335	0.3089	0.3989	0.4064	0.1374
LARCM + Novelty	0.2991	0.3826	0.3894	0.1143	0.3089	0.3904	0.4009	0.1188
LARCM + Support	0.2991	0.3702	0.4046	**0.1641***	0.3089	0.3933	0.4150	**0.1660***
LDA + *bag-of-words*	0.2991	0.2979	0.3053	0.0378	0.3089	0.3076	0.3150	0.0394
LDA + *bag-of-related-words*	0.2991	0.3780	0.3903	0.1138	0.3089	0.3933	0.4017	0.1178

5.4. Experimental setup

To measure the predictive ability of the recommender systems, we used the All But One protocol (Breese, Heckerman, & Kadie, 1998) with 10-fold cross-validation, and calculated the mean average precision (MAP). To do this, the sessions in the dataset were randomly partitioned into 10 subsets. For each fold, we used 9 of those subsets of data for training, and the remaining one for testing. The training set T_r was used to build the recommendation model. For each user in the test set T_e, we randomly hid one item, referred to as the singleton set H. The remaining items represented the set of observables, O, on which the recommendation was based. Finally, we computed the MAP, as follows.

The mean average precision metric computes the precision, considering the respective position in the ordered list of recommended items. From this measure, we obtain a single accuracy score for a set of test users T_e:

$$MAP(T_e) = \frac{1}{|T_e|}\sum_{j=1}^{|T_e|} AveP(R_j, H_j), \tag{8}$$

where the average precision (AveP) is given by

$$AveP(R_j, H_j) = \frac{1}{|H_j|}\sum_{r=1}^{|H_j|}[Prec(R_j, r) \times \delta(R_j(r), H_j)], \tag{9}$$

where $Prec(R_j, r)$ is the precision for all recommended items up to ranking r and $\delta(R_j(r), H_j) = 1$, if the predicted item at ranking r is a relevant item ($R_j(r) \in H_j$) or otherwise zero.

We computed MAP@N, for N equal to 5 and 10 recommendations. The notation @ refers to the top ranked recommendations. For each combination and measure, we calculate the mean to summarize the 10-fold values. To compare two recommendation algorithms, we applied the two-sided paired t-test with a 95% confidence level (Mitchell, 1997). To run the evaluation, we used the CARSlibrary[13], a library with context-aware recommender algorithms for Top-N recommendations.

5.5. Results

The values of MAP@5 and MAP@10 for each LARCM and LDA model are presented in Tables 10–12. The best results for each recommendation algorithm are marked with an asterisk ("*"). By examining the tables, we can verify that, considering the topics extracted by our proposal, the algorithms cReduction and weightPoF presented statistically better results in relation to the baseline IBCF for almost all the cases. Concerning the algorithm filterPoF, the results were not so stable with the parameter variations of the topic models. We can also observe that the results obtained by the LARCM were superior to those obtained by the LDA model.

Regarding the results obtained with the proposed LARCM, the recommender algorithms presented the best results for the objective measure ϕ-Coefficient and for the value of k equal to 150. Note that the measure ϕ-Coefficient also presented the best results in the topic interpretability evaluation, and was very competitive in the document classification evaluation. These results show that the measure ϕ-Coefficient is a good candidate for generating topics with the LARCM. The Support measure, which presented the best results for the classification task, did not show significant results in the context-aware recommender systems. Taking a look at the results for each value of k, when k is equal to 50, the measure Support presented the best results for the algorithm filterPoF, and the Kappa measure presented the best results for the other algorithms. For k equal to 100, the measure Mutual Information LHS presented the best value for all algorithms. In both cases of k being equal to 50 and 100, the measure ϕ-Coefficient presented results that approximated the best results presented by the algorithms cReduction and weightPoF. For k equal to 150, the measure ϕ-Coefficient presented the best value for all algorithms. In general, we can see that the contextual information provided by the LARCM can contribute to generating better recommendations.

6. Conclusion and future work

In this work we proposed the latent association rule cluster based model (LARCM). This is a non-probabilistic topic modeling approach that makes use of clustering of association rules to provide a document representation with correlation of terms and low

[13] https://github.com/maddomingues/CARSlibrary.

Table 11
Comparing the context-aware recommenders against the IBCF algorithm. The contextual information (i.e., topics) was obtained by using the LARCM and the LDA model, while considering the value of k equal to 100.

	MAP @ 5				MAP @ 10			
	IBCF	cReduction	weightPoF	filterPoF	IBCF	cReduction	weightPoF	filterPoF
LARCM + ϕ-Coefficient	0.2991	0.4752	0.4899	0.1699	0.3089	0.4821	0.4962	0.1742
LARCM + Added Value	0.2991	0.4302	0.4749	0.3892	0.3089	0.4379	0.4789	0.3892
LARCM + Certainty Factor	0.2991	0.3724	0.4312	0.4066	0.3089	0.3912	0.4433	0.4069
LARCM + Collective Strength	0.2991	0.4908	0.5044	0.1889	0.3089	0.4939	0.5067	0.1950
LARCM + Confidence	0.2991	0.4231	0.4580	0.4013	0.3089	0.4328	0.4612	0.4033
LARCM + Conviction	0.2991	0.3740	0.3812	0.1199	0.3089	0.3890	0.3965	0.1247
LARCM + Gini Index	0.2991	0.5045	0.5199	0.3881	0.3089	0.5093	0.5297	0.3911
LARCM + IS	0.2991	0.3530	0.4163	0.2606	0.3089	0.3776	0.4288	0.2633
LARCM + J-Measure	0.2991	0.3940	0.4348	0.2195	0.3089	0.4173	0.4482	0.2201
LARCM + Kappa	0.2991	0.4845	0.5038	0.2721	0.3089	0.4916	0.5088	0.2762
LARCM + Klosgen	0.2991	0.3851	0.4132	0.1792	0.3089	0.4033	0.4250	0.1822
LARCM + Lambda	0.2991	0.4909	0.5270	0.3671	0.3089	0.4923	0.5291	0.3704
LARCM + Laplace	0.2991	0.4121	0.4277	0.1459	0.3089	0.4219	0.4372	0.1493
LARCM + Lift	0.2991	0.4269	0.4721	0.3595	0.3089	0.4351	0.4798	0.3619
LARCM + Mutual Information	0.2991	**0.5085***	**0.5301***	**0.4497***	0.3089	**0.5112***	**0.5325***	**0.4522***
LARCM + Novelty	0.2991	0.3934	0.4119	0.2315	0.3089	0.4015	0.4238	0.2342
LARCM + Support	0.2991	0.0000	0.4429	0.2236	0.3089	0.0000	0.4493	0.2257
LDA + *bag-of-words*	0.2991	0.2987	0.3076	0.0477	0.3089	0.3014	0.3176	0.0492
LDA + *bag-of-related-words*	0.2991	0.3783	0.3874	0.1193	0.3089	0.3889	0.3987	0.1234

Table 12
Comparing the context-aware recommenders against the IBCF algorithm. The contextual information (i.e., topics) was obtained by using the LARCM and the LDA model, while considering the value of k equal to 150.

	MAP @ 5				MAP @ 10			
	IBCF	cReduction	weightPoF	filterPoF	IBCF	cReduction	weightPoF	filterPoF
LARCM + ϕ-Coefficient	0.2991	**0.6309***	**0.6506***	**0.6115***	0.3089	**0.6324***	**0.6522***	**0.6120***
LARCM + Added Value	0.2991	0.4890	0.5147	0.3991	0.3089	0.4777	0.5184	0.4003
LARCM + Certainty Factor	0.2991	0.0000	0.4703	0.4170	0.3089	0.0000	0.4744	0.4206
LARCM + Collective Strength	0.2991	0.4897	0.5324	0.4088	0.3089	0.4927	0.5360	0.4097
LARCM + Confidence	0.2991	0.0000	0.5584	0.4420	0.3089	0.0000	0.5619	0.4427
LARCM + Conviction	0.2991	0.3562	0.3626	0.1060	0.3089	0.3713	0.3779	0.1106
LARCM + Gini Index	0.2991	0.5269	0.5450	0.5081	0.3089	0.5332	0.5469	0.5091
LARCM + IS	0.2991	0.4846	0.5446	0.2854	0.3089	0.4902	0.5487	0.2892
LARCM + J-Measure	0.2991	0.0000	0.6486	0.5274	0.3089	0.0000	0.6498	0.5282
LARCM + Kappa	0.2991	0.5903	0.6037	0.5753	0.3089	0.5928	0.6062	0.5758
LARCM + Klosgen	0.2991	0.0000	0.5722	0.4209	0.3089	0.0000	0.5748	0.4209
LARCM + Lambda	0.2991	0.0000	0.5024	0.4045	0.3089	0.0000	0.5078	0.4081
LARCM + Laplace	0.2991	0.0000	0.6405	0.3752	0.3089	0.0000	0.6432	0.3768
LARCM + Lift	0.2991	0.3746	0.4504	0.2565	0.3089	0.3729	0.4519	0.2605
LARCM + Mutual Information	0.2991	0.5286	0.5697	0.4992	0.3089	0.5336	0.5725	0.5015
LARCM + Novelty	0.2991	0.5208	0.5544	0.2522	0.3089	0.5313	0.5580	0.2535
LARCM + Support	0.2991	0.3846	0.4461	0.3109	0.3089	0.3973	0.4517	0.3122
LDA + *bag-of-words*	0.2991	0.3046	0.3117	0.0563	0.3089	0.3151	0.3218	0.0577
LDA + *bag-of-related-words*	0.2991	0.3687	0.3860	0.1181	0.3089	0.3868	0.3965	0.1222

dimensionality. In our proposal, association rules are built for each document to extract the correlated terms. We called these relations between terms the *local context* of relations. Then, we apply a clustering algorithm to all association rules in order to discover the *general context* of relations. Each cluster becomes a topic for the new document representation. The main idea behind our proposal is that the neighborhood of correlated terms will reveal different terms that are either used in the same context or have the same meaning, and identical terms that are either used in different contexts or have different meanings.

We evaluated the interpretability of the topics obtained with the LARCM against the ones produced by the traditional LDA model and by the LDA model using a document representation that includes correlated terms (i.e., bag-of-related-words). The LDA model is state-of-the-art for topic extraction models, yet experimental results demonstrated that the LARCM provides topics with better interpretability than both LDA models. Additionally, we used the topics obtained by the LARCM with two different applications: text classification and page recommendation. Regarding text classification, the topics were used to improve document collection representation. As for page recommendation, the topics were used as contextual information in context-aware recommender systems. Results have shown that the topics provided by the LARCM can be used to improve both applications.

As future work we can mention:

- The replication of the experiments in other non-textual scenarios, such as image processing and time series. Although our focus in this work was to deal with text-based information, the topic models in the literature are also applied to other (i.e., non-textual) scenarios;
- An extension of experiments with neural probabilistic language models such as word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and doc2vec (Le & Mikolov, 2014). They are unsupervised learning algorithms that learn vector representations by using neural network models. The vector representations are learned based on the words in contexts by using a strategy for dimensionality reduction that creates new features;
- Introducing variation to the objective measure used at two distinct points during the running of the LARCM, i.e., the selection

of the association rules and the set of topic descriptors. In our work, we used the same measure in both moments, but the LARCM allows us to use different measures for each;
- Since the association rules extraction used by LARCM is able to process each document individually, we intend to re-implement our proposal in a parallel and distributed way, by using the MapReduce paradigm, so that we can apply our proposal to big data collections.

Acknowledgment

This work is supported by Araucaria Foundation (Paraná/Brazil), CAPES/Brazil and CNPq/Brazil.

Appendix A. Results for the different classifiers and representation models in the classification task

In this appendix, we present the accuracy and standard deviation for each classifier and representation model used in the classification task. As already stated in Section 5.1, the classification algorithms are: Naive Bayes (NB); Multinomial Naive Bayes (MNB); J48 (J48); SMO (SMO) using the generalization parameter "c" with its value equal to 1 and 10; e IBk (IBk) with the $\mathbb{k}$ neighbors defined in 3, 5 and 7. (Tables A1–A27)

Table A1
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-1.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	78.42 ± 4.93	79.48 ± 5.79	77.48 ± 6.63	76.97 ± 5.62	80.33 ± 5.69	68.35 ± 6.21	69.50 ± 6.80	68.25 ± 6.88
LDA + *bag-of-words*	86.04 ± 4.98	84.77 ± 5.44	82.03 ± 5.06	85.57 ± 5.02	84.29 ± 5.01	80.05 ± 6.39	81.25 ± 6.18	81.10 ± 6.50
LARCM + Added Value	73.18 ± 6.51	74.39 ± 5.73	66.99 ± 6.98	74.26 ± 6.08	77.09 ± 5.72	66.98 ± 6.74	68.71 ± 7.08	69.37 ± 6.71
LARCM + Certainty Factor	71.03 ± 6.66	74.56 ± 6.54	71.33 ± 6.58	72.50 ± 6.93	73.33 ± 6.36	62.98 ± 7.19	64.98 ± 6.80	63.93 ± 7.24
LARCM + Collective Strength	75.08 ± 7.23	75.72 ± 7.10	69.69 ± 6.56	72.79 ± 7.07	75.69 ± 6.75	65.89 ± 7.11	64.74 ± 6.96	67.77 ± 6.58
LARCM + Confidence	75.29 ± 6.37	78.02 ± 6.14	66.11 ± 7.56	75.64 ± 6.12	78.47 ± 5.84	65.24 ± 6.78	66.59 ± 7.41	66.87 ± 7.50
LARCM + Conviction	55.95 ± 8.19	60.70 ± 7.99	55.90 ± 8.11	60.03 ± 7.45	60.08 ± 7.57	50.83 ± 7.11	53.69 ± 7.50	53.82 ± 7.58
LARCM + ϕ-Coefficient	71.71 ± 6.81	75.34 ± 6.60	67.70 ± 7.05	72.61 ± 6.73	77.27 ± 6.28	65.40 ± 5.97	64.99 ± 7.37	65.59 ± 7.09
LARCM + Gini Index	73.24 ± 6.50	74.60 ± 7.17	66.45 ± 7.29	72.40 ± 7.29	75.35 ± 6.57	66.12 ± 6.39	66.05 ± 6.31	67.51 ± 6.49
LARCM + IS	75.44 ± 6.02	75.14 ± 5.36	70.43 ± 6.04	74.53 ± 6.76	76.24 ± 6.28	64.03 ± 7.34	64.74 ± 6.96	65.41 ± 6.68
LARCM + J-Measure	72.86 ± 6.43	72.96 ± 6.41	64.31 ± 7.18	71.20 ± 6.70	74.94 ± 6.28	64.51 ± 7.09	65.14 ± 6.35	65.31 ± 6.78
LARCM + Kappa	74.41 ± 7.56	75.79 ± 6.14	69.38 ± 6.02	74.37 ± 6.21	79.27 ± 6.31	64.79 ± 6.70	67.33 ± 6.59	65.90 ± 6.98
LARCM + Klosgen	71.65 ± 6.86	75.34 ± 6.10	64.35 ± 7.54	73.07 ± 7.06	75.74 ± 7.08	63.20 ± 6.95	64.53 ± 7.19	66.53 ± 7.49
LARCM + Lambda	67.11 ± 6.45	74.60 ± 5.84	61.43 ± 7.10	74.28 ± 5.96	75.73 ± 5.93	64.00 ± 7.57	64.88 ± 7.37	66.21 ± 7.43
LARCM + Laplace	74.27 ± 6.29	76.62 ± 5.26	69.35 ± 6.94	75.39 ± 5.05	75.54 ± 5.90	64.85 ± 5.87	66.02 ± 6.22	66.77 ± 6.16
LARCM + Lift	72.08 ± 5.25	73.48 ± 5.97	65.44 ± 7.27	72.66 ± 5.83	76.24 ± 6.21	61.53 ± 6.54	64.33 ± 6.08	65.36 ± 6.42
LARCM + Mutual Information LHS	69.82 ± 6.94	72.92 ± 6.51	66.81 ± 7.17	71.12 ± 6.52	74.53 ± 6.92	62.65 ± 6.49	64.75 ± 6.98	64.93 ± 6.35
LARCM + Novelty	71.94 ± 6.93	77.12 ± 6.65	69.15 ± 7.08	75.46 ± 6.60	76.29 ± 6.52	67.12 ± 6.92	68.87 ± 6.53	68.87 ± 6.39
LARCM + Support	79.40 ± 6.57	79.97 ± 6.22	75.92 ± 6.95	79.26 ± 6.80	82.27 ± 6.38	72.37 ± 6.93	73.95 ± 7.51	73.65 ± 7.78

Table A2
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-1.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	77.07 ± 6.88	73.64 ± 7.52	69.56 ± 6.54	75.16 ± 6.98	78.25 ± 6.54	64.59 ± 7.06	67.55 ± 6.83	71.74 ± 7.26
LDA + *bag-of-words*	78.55 ± 6.46	80.03 ± 5.76	72.71 ± 7.04	77.47 ± 6.27	79.80 ± 6.28	72.09 ± 7.16	72.46 ± 6.71	71.88 ± 6.64
LARCM + Added Value	76.09 ± 6.10	73.53 ± 6.13	63.84 ± 6.11	73.55 ± 5.93	75.71 ± 6.65	61.60 ± 6.77	62.91 ± 7.55	63.48 ± 7.59
LARCM + Certainty Factor	78.90 ± 6.30	77.29 ± 6.52	69.94 ± 6.20	74.84 ± 6.54	76.57 ± 6.98	65.42 ± 6.53	68.35 ± 6.57	70.06 ± 6.27
LARCM + Collective Strength	77.07 ± 6.41	74.11 ± 6.87	67.89 ± 7.14	72.56 ± 7.57	75.49 ± 6.18	61.15 ± 6.82	60.60 ± 7.53	62.58 ± 7.57
LARCM + Confidence	80.33 ± 5.79	75.19 ± 6.30	73.43 ± 7.00	74.09 ± 5.97	75.21 ± 6.24	59.67 ± 7.46	62.66 ± 7.02	63.47 ± 6.05
LARCM + Conviction	54.61 ± 7.15	59.25 ± 7.88	55.64 ± 7.80	57.17 ± 7.39	59.17 ± 7.44	44.76 ± 7.52	46.87 ± 7.56	48.21 ± 8.12
LARCM + ϕ-Coefficient	79.07 ± 6.57	77.27 ± 6.76	72.28 ± 7.33	75.14 ± 6.93	78.27 ± 7.03	67.10 ± 7.38	66.60 ± 6.83	68.12 ± 6.64
LARCM + Gini Index	77.94 ± 6.61	75.63 ± 6.44	71.27 ± 7.44	74.71 ± 6.83	77.64 ± 7.06	67.09 ± 6.45	68.42 ± 6.97	69.60 ± 7.39
LARCM + IS	76.06 ± 7.26	74.64 ± 5.82	71.25 ± 6.83	73.98 ± 6.27	76.21 ± 6.42	60.81 ± 7.03	61.95 ± 7.25	63.33 ± 6.62
LARCM + J-Measure	76.85 ± 6.34	76.13 ± 6.41	69.22 ± 6.82	73.82 ± 5.94	76.98 ± 5.51	63.83 ± 7.46	64.84 ± 6.81	66.42 ± 7.07
LARCM + Kappa	75.92 ± 6.87	74.37 ± 6.25	65.63 ± 6.55	72.55 ± 6.92	74.08 ± 5.79	64.36 ± 6.13	65.31 ± 6.78	64.94 ± 6.78
LARCM + Klosgen	78.37 ± 6.07	73.33 ± 6.62	72.46 ± 6.93	71.43 ± 6.74	75.74 ± 6.66	63.25 ± 7.07	64.10 ± 7.19	63.88 ± 6.48
LARCM + Lambda	72.25 ± 7.71	71.62 ± 6.61	66.97 ± 7.60	71.12 ± 7.57	73.83 ± 6.08	59.63 ± 7.75	61.55 ± 7.88	62.60 ± 7.61
LARCM + Laplace	78.32 ± 6.76	76.04 ± 7.00	71.75 ± 7.00	74.14 ± 7.28	77.40 ± 7.20	64.51 ± 7.54	66.59 ± 6.38	67.74 ± 7.83
LARCM + Lift	74.81 ± 5.73	74.29 ± 6.19	68.04 ± 6.89	74.14 ± 6.28	76.54 ± 6.45	59.44 ± 6.94	60.50 ± 6.71	62.28 ± 6.04
LARCM + Mutual Information LHS	78.25 ± 6.40	74.48 ± 6.60	72.63 ± 7.43	72.86 ± 6.91	75.57 ± 6.43	62.90 ± 7.19	67.19 ± 7.61	68.12 ± 7.83
LARCM + Novelty	78.97 ± 5.77	74.91 ± 5.66	74.36 ± 6.74	74.31 ± 5.85	75.39 ± 6.15	64.01 ± 6.49	66.84 ± 5.78	68.07 ± 6.15
LARCM + Support	80.85 ± 4.91	79.76 ± 6.02	74.81 ± 6.31	78.31 ± 6.17	84.03 ± 5.01	74.28 ± 6.41	73.25 ± 6.59	75.20 ± 5.77

Table A3
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-1.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	73.81 ± 6.34	66.54 ± 7.38	62.59 ± 6.81	69.40 ± 6.82	70.63 ± 6.72	61.41 ± 7.32	64.82 ± 6.78	67.05 ± 7.07
LDA + *bag-of-words*	71.54 ± 6.86	71.36 ± 6.35	62.34 ± 8.02	53.09 ± 6.36	79.58 ± 6.23	68.90 ± 7.32	64.74 ± 7.55	62.43 ± 9.50
LARCM + Added Value	77.61 ± 6.23	74.36 ± 6.60	66.59 ± 7.34	70.58 ± 5.90	75.36 ± 6.86	59.16 ± 7.54	63.10 ± 7.48	63.70 ± 6.80
LARCM + Certainty Factor	78.64 ± 5.99	73.93 ± 6.85	67.18 ± 6.67	71.06 ± 7.31	76.13 ± 6.29	63.06 ± 6.66	64.55 ± 6.83	64.57 ± 7.57
LARCM + Collective Strength	80.88 ± 6.34	72.69 ± 7.07	66.35 ± 7.47	70.21 ± 7.38	77.95 ± 7.29	61.07 ± 7.05	61.96 ± 7.95	62.68 ± 8.02
LARCM + Confidence	77.36 ± 6.11	72.93 ± 5.92	72.83 ± 7.45	70.23 ± 7.72	75.06 ± 7.20	64.63 ± 7.50	65.47 ± 8.09	67.68 ± 7.20
LARCM + Conviction	57.25 ± 7.65	60.40 ± 6.62	53.41 ± 7.63	58.18 ± 7.30	60.55 ± 7.16	45.52 ± 7.93	47.24 ± 8.20	49.78 ± 7.83
LARCM + ϕ-Coefficient	79.50 ± 6.95	75.51 ± 6.37	70.38 ± 6.37	72.25 ± 6.81	78.51 ± 5.24	60.84 ± 6.51	66.58 ± 6.23	67.77 ± 6.37
LARCM + Gini Index	77.74 ± 6.25	75.84 ± 5.96	70.31 ± 7.05	72.80 ± 6.44	78.27 ± 6.00	62.46 ± 7.33	66.89 ± 6.18	68.19 ± 7.19
LARCM + IS	78.75 ± 7.06	74.09 ± 5.59	67.52 ± 6.78	72.11 ± 5.98	76.40 ± 6.14	62.14 ± 6.82	64.09 ± 6.89	68.45 ± 6.50
LARCM + J-Measure	80.10 ± 6.07	74.49 ± 6.78	70.49 ± 7.08	72.51 ± 6.72	78.20 ± 5.64	61.67 ± 7.19	63.41 ± 7.68	65.34 ± 6.77
LARCM + Kappa	78.89 ± 5.90	75.33 ± 6.59	71.48 ± 6.63	72.57 ± 6.65	76.67 ± 6.92	65.03 ± 6.99	66.83 ± 6.57	68.66 ± 6.54
LARCM + Klosgen	78.34 ± 6.42	76.00 ± 6.20	66.89 ± 6.64	73.38 ± 6.25	76.41 ± 6.14	63.90 ± 6.42	64.99 ± 6.86	68.62 ± 6.46
LARCM + Lambda	74.19 ± 6.96	72.88 ± 5.24	63.16 ± 7.77	69.35 ± 5.65	75.81 ± 6.33	61.26 ± 7.01	62.99 ± 6.75	64.34 ± 7.05
LARCM + Laplace	79.72 ± 6.05	71.40 ± 6.83	71.05 ± 7.19	70.90 ± 6.67	75.14 ± 6.50	57.97 ± 6.98	61.17 ± 6.93	62.35 ± 6.88
LARCM + Lift	76.98 ± 6.27	69.34 ± 6.69	66.01 ± 7.20	67.86 ± 7.13	75.93 ± 6.79	61.84 ± 8.62	62.53 ± 8.01	64.17 ± 8.40
LARCM + Mutual Information LHS	78.09 ± 6.72	73.24 ± 7.22	64.97 ± 8.00	72.76 ± 7.00	76.27 ± 6.21	60.57 ± 7.45	63.89 ± 6.74	66.16 ± 6.49
LARCM + Novelty	76.28 ± 6.98	74.21 ± 6.55	70.32 ± 6.54	71.65 ± 6.44	76.66 ± 6.34	63.91 ± 6.90	64.94 ± 7.62	66.59 ± 7.52
LARCM + Support	82.73 ± 5.92	78.62 ± 6.26	79.23 ± 6.59	75.08 ± 7.11	82.73 ± 6.04	67.92 ± 5.89	71.40 ± 6.76	73.98 ± 6.72

Table A4
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-2.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	87.78 ± 4.85	88.61 ± 4.79	85.39 ± 5.43	84.41 ± 5.36	87.17 ± 5.20	77.83 ± 6.36	77.22 ± 6.47	79.68 ± 6.02
LDA + *bag-of-words*	94.93 ± 3.09	95.29 ± 3.01	85.83 ± 5.17	94.00 ± 3.33	94.37 ± 3.23	93.49 ± 3.36	92.98 ± 3.77	93.02 ± 3.54
LARCM + Added Value	83.15 ± 5.33	89.00 ± 4.81	82.88 ± 5.63	88.29 ± 4.48	89.73 ± 4.21	79.80 ± 5.57	81.61 ± 5.57	82.85 ± 5.49
LARCM + Certainty Factor	82.73 ± 6.05	89.85 ± 4.26	82.93 ± 5.83	88.61 ± 4.72	88.90 ± 4.59	77.15 ± 6.09	80.66 ± 6.10	81.76 ± 5.40
LARCM + Collective Strength	82.15 ± 5.01	89.44 ± 4.44	79.51 ± 5.44	89.10 ± 4.57	90.05 ± 4.42	78.73 ± 5.84	79.20 ± 5.93	79.90 ± 5.64
LARCM + Confidence	81.61 ± 5.54	87.27 ± 4.58	82.17 ± 5.48	87.90 ± 4.41	88.00 ± 4.70	78.71 ± 5.18	80.34 ± 5.60	80.44 ± 5.00
LARCM + Conviction	70.98 ± 7.07	74.39 ± 6.74	72.41 ± 6.66	73.49 ± 7.08	72.61 ± 6.70	67.12 ± 6.81	68.61 ± 7.42	70.07 ± 7.21
LARCM + ϕ-Coefficient	81.66 ± 5.18	90.44 ± 4.21	81.88 ± 5.43	90.56 ± 4.02	91.07 ± 3.56	83.37 ± 5.67	83.88 ± 5.11	84.98 ± 5.04
LARCM + Gini Index	84.39 ± 5.52	88.98 ± 4.90	79.46 ± 5.61	88.85 ± 4.55	89.56 ± 5.00	80.37 ± 6.29	81.66 ± 6.63	81.76 ± 6.15
LARCM + IS	81.49 ± 5.41	87.88 ± 4.78	80.12 ± 6.65	87.34 ± 5.25	89.20 ± 4.05	79.80 ± 4.99	81.63 ± 5.48	81.73 ± 5.34
LARCM + J-Measure	83.29 ± 5.52	88.88 ± 4.63	81.61 ± 5.61	89.07 ± 4.96	89.20 ± 4.91	80.32 ± 5.93	82.49 ± 5.58	83.85 ± 5.15
LARCM + Kappa	84.39 ± 5.93	89.15 ± 4.14	79.98 ± 6.27	89.59 ± 3.94	90.24 ± 4.32	80.24 ± 5.52	82.63 ± 5.54	82.44 ± 5.45
LARCM + Klosgen	84.12 ± 5.36	87.90 ± 4.99	81.20 ± 5.39	87.32 ± 4.77	89.78 ± 5.10	80.68 ± 5.98	80.51 ± 5.83	79.71 ± 5.79
LARCM + Lambda	82.17 ± 6.03	86.68 ± 4.85	79.17 ± 6.47	87.07 ± 4.90	86.46 ± 5.19	78.12 ± 6.28	79.66 ± 5.90	80.02 ± 6.40
LARCM + Laplace	83.22 ± 6.23	87.02 ± 4.97	79.61 ± 6.10	86.83 ± 5.29	87.61 ± 5.08	79.66 ± 6.23	80.68 ± 6.23	80.85 ± 6.40
LARCM + Lift	80.44 ± 5.85	88.44 ± 4.82	77.10 ± 6.24	87.29 ± 5.00	87.37 ± 4.49	77.24 ± 6.59	78.39 ± 6.83	79.51 ± 6.35
LARCM + Mutual Information LHS	82.66 ± 5.40	87.44 ± 4.78	81.51 ± 5.25	86.73 ± 4.77	87.49 ± 4.95	79.05 ± 5.79	81.20 ± 5.37	81.85 ± 5.31
LARCM + Novelty	82.66 ± 6.07	87.46 ± 4.87	84.39 ± 5.54	86.63 ± 5.10	88.32 ± 4.48	79.15 ± 6.01	81.32 ± 5.81	81.68 ± 5.48
LARCM + Support	88.68 ± 4.81	91.78 ± 4.38	82.61 ± 5.96	91.95 ± 4.02	92.68 ± 4.33	86.63 ± 5.58	86.20 ± 5.85	86.20 ± 5.65

Table A5
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-2.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	86.95 ± 4.80	85.02 ± 5.13	78.32 ± 6.98	83.68 ± 5.32	86.29 ± 4.92	68.93 ± 7.41	70.17 ± 7.51	76.51 ± 6.63
LDA + *bag-of-words*	88.05 ± 4.38	88.15 ± 4.70	79.15 ± 5.75	86.15 ± 4.95	86.66 ± 4.98	79.46 ± 5.95	79.83 ± 6.22	77.41 ± 5.88
LARCM + Added Value	85.17 ± 4.65	89.24 ± 3.83	82.27 ± 6.17	88.93 ± 3.63	89.88 ± 4.11	77.46 ± 5.97	80.12 ± 5.57	81.56 ± 5.26
LARCM + Certainty Factor	84.78 ± 5.03	88.15 ± 5.18	87.41 ± 6.47	87.63 ± 5.54	89.68 ± 4.69	77.27 ± 6.70	79.22 ± 6.21	80.49 ± 5.79
LARCM + Collective Strength	81.34 ± 4.75	88.71 ± 5.17	79.68 ± 5.55	88.27 ± 4.70	90.20 ± 4.23	76.20 ± 6.01	80.37 ± 6.01	80.27 ± 6.00
LARCM + Confidence	85.46 ± 4.88	87.93 ± 4.81	82.71 ± 5.75	87.95 ± 4.52	91.07 ± 3.82	80.63 ± 5.66	83.05 ± 5.45	84.10 ± 5.26
LARCM + Conviction	72.71 ± 7.64	75.39 ± 6.71	67.76 ± 7.02	74.78 ± 6.83	74.29 ± 6.66	64.54 ± 7.14	67.10 ± 7.33	65.32 ± 7.69
LARCM + ϕ-Coefficient	84.71 ± 5.05	88.88 ± 4.02	82.34 ± 5.03	89.95 ± 3.45	91.44 ± 3.81	79.49 ± 6.24	83.66 ± 5.69	85.68 ± 5.03
LARCM + Gini Index	85.63 ± 5.07	88.80 ± 4.19	82.34 ± 5.46	88.66 ± 3.99	89.90 ± 4.23	77.27 ± 6.18	79.90 ± 5.26	82.39 ± 4.92
LARCM + IS	84.68 ± 5.45	89.02 ± 4.92	83.29 ± 5.41	88.24 ± 5.19	88.71 ± 4.30	78.46 ± 6.31	79.54 ± 6.49	81.24 ± 6.01
LARCM + J-Measure	85.39 ± 5.10	88.39 ± 4.63	82.27 ± 6.44	87.27 ± 4.82	88.10 ± 4.40	77.71 ± 6.30	80.71 ± 6.50	80.56 ± 6.19
LARCM + Kappa	84.10 ± 5.89	87.29 ± 5.28	85.00 ± 5.40	87.68 ± 5.20	88.98 ± 4.34	79.02 ± 6.39	81.54 ± 6.11	82.37 ± 6.06
LARCM + Klosgen	81.83 ± 6.30	88.76 ± 4.17	87.24 ± 6.21	87.17 ± 4.47	88.78 ± 4.03	79.93 ± 5.52	81.05 ± 5.82	80.29 ± 5.47
LARCM + Lambda	83.24 ± 4.89	87.17 ± 4.77	79.80 ± 5.61	86.54 ± 5.02	88.22 ± 5.02	77.34 ± 5.65	77.85 ± 6.37	78.41 ± 6.23
LARCM + Laplace	82.78 ± 5.48	88.10 ± 4.77	82.78 ± 5.76	86.66 ± 5.05	88.02 ± 4.99	78.34 ± 6.54	79.85 ± 6.36	81.54 ± 6.45
LARCM + Lift	84.83 ± 4.96	88.76 ± 5.39	81.22 ± 5.42	87.88 ± 5.23	88.85 ± 5.18	78.34 ± 6.17	80.12 ± 6.73	82.56 ± 6.79
LARCM + Mutual Information LHS	83.76 ± 5.70	88.88 ± 4.64	82.49 ± 5.47	87.78 ± 5.19	89.71 ± 4.53	79.15 ± 5.45	81.39 ± 5.22	81.56 ± 4.76
LARCM + Novelty	84.61 ± 5.09	86.59 ± 5.61	81.07 ± 5.84	86.56 ± 5.71	88.49 ± 4.92	79.17 ± 5.53	81.46 ± 5.56	81.44 ± 5.57
LARCM + Support	89.83 ± 5.11	91.22 ± 4.30	86.85 ± 5.01	91.07 ± 4.36	92.46 ± 4.01	82.68 ± 5.86	85.54 ± 5.57	85.59 ± 5.84

Table A6
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-2.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	79.51 ± 5.95	81.27 ± 5.55	65.00 ± 7.31	80.02 ± 5.55	81.51 ± 5.55	70.05 ± 6.63	66.46 ± 7.72	61.63 ± 7.07
LDA + *bag-of-words*	78.76 ± 5.43	81.80 ± 4.67	69.51 ± 6.80	79.56 ± 5.41	82.98 ± 5.55	71.41 ± 6.30	73.07 ± 6.33	67.95 ± 6.86
LARCM + Added Value	86.02 ± 5.13	89.12 ± 4.57	82.32 ± 5.73	87.73 ± 4.88	90.49 ± 4.17	79.17 ± 5.71	81.93 ± 5.87	83.24 ± 5.71
LARCM + Certainty Factor	86.61 ± 5.24	88.17 ± 5.15	87.22 ± 4.81	86.78 ± 5.25	88.80 ± 4.47	79.24 ± 5.91	82.29 ± 5.47	83.05 ± 5.58
LARCM + Collective Strength	85.39 ± 5.58	87.07 ± 5.27	83.07 ± 5.87	84.78 ± 5.60	87.07 ± 5.70	76.93 ± 6.12	78.76 ± 6.22	80.27 ± 6.02
LARCM + Confidence	83.66 ± 6.23	87.80 ± 4.84	83.66 ± 5.53	86.22 ± 5.09	90.10 ± 4.71	76.22 ± 5.60	78.20 ± 6.69	79.12 ± 5.99
LARCM + Conviction	67.59 ± 6.93	74.29 ± 6.36	65.85 ± 7.23	72.66 ± 5.91	72.95 ± 6.73	62.93 ± 7.65	65.10 ± 6.67	66.32 ± 6.71
LARCM + ϕ-Coefficient	87.10 ± 4.71	88.98 ± 5.11	83.66 ± 5.73	88.34 ± 5.14	90.07 ± 4.36	78.41 ± 5.59	80.56 ± 5.73	81.51 ± 5.18
LARCM + Gini Index	88.10 ± 5.31	88.88 ± 4.80	85.46 ± 5.08	86.80 ± 4.98	90.10 ± 4.21	79.12 ± 5.86	79.85 ± 5.29	82.07 ± 5.66
LARCM + IS	85.41 ± 4.62	87.10 ± 4.18	86.80 ± 4.92	85.24 ± 5.06	87.63 ± 4.60	77.61 ± 6.21	78.61 ± 6.32	79.80 ± 6.19
LARCM + J-Measure	85.37 ± 5.63	88.05 ± 5.25	84.98 ± 5.84	86.24 ± 5.15	88.44 ± 4.68	77.07 ± 6.28	79.49 ± 6.32	81.22 ± 6.42
LARCM + Kappa	83.56 ± 5.46	89.05 ± 4.04	83.12 ± 5.72	87.71 ± 4.76	90.02 ± 4.41	77.46 ± 5.19	81.39 ± 5.59	82.66 ± 5.52
LARCM + Klosgen	85.46 ± 5.37	88.44 ± 4.51	83.95 ± 5.14	87.93 ± 4.87	89.66 ± 4.23	76.85 ± 6.02	79.22 ± 6.07	80.85 ± 5.89
LARCM + Lambda	86.20 ± 4.88	87.27 ± 4.87	78.90 ± 5.93	85.15 ± 5.04	85.32 ± 5.85	76.05 ± 6.81	77.88 ± 6.22	77.85 ± 6.08
LARCM + Laplace	85.22 ± 5.06	87.93 ± 4.78	85.68 ± 5.18	85.63 ± 5.11	87.73 ± 4.32	79.07 ± 5.34	82.15 ± 4.97	83.02 ± 5.08
LARCM + Lift	84.61 ± 6.07	88.85 ± 4.94	80.95 ± 6.32	86.44 ± 4.96	88.78 ± 4.40	76.00 ± 6.63	78.95 ± 5.93	79.78 ± 6.23
LARCM + Mutual Information LHS	84.78 ± 5.41	88.37 ± 5.02	81.41 ± 5.47	87.39 ± 4.98	87.00 ± 5.01	80.27 ± 7.21	83.73 ± 6.28	83.76 ± 6.07
LARCM + Novelty	86.00 ± 4.67	87.76 ± 4.62	82.59 ± 5.89	87.02 ± 4.75	90.49 ± 4.34	80.85 ± 6.30	82.56 ± 5.76	84.15 ± 5.48
LARCM + Support	90.93 ± 4.44	91.20 ± 4.20	86.07 ± 5.41	90.76 ± 4.14	93.00 ± 3.78	85.49 ± 5.39	87.12 ± 4.77	88.39 ± 4.53

Table A7
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-3.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	86.39 ± 4.83	84.66 ± 4.88	84.37 ± 4.99	80.02 ± 5.41	83.01 ± 5.02	73.86 ± 6.40	74.90 ± 5.72	77.45 ± 5.64
LDA + *bag-of-words*	91.19 ± 4.17	92.07 ± 4.04	89.20 ± 4.69	91.73 ± 4.22	91.20 ± 4.29	87.12 ± 4.47	89.02 ± 4.73	88.48 ± 4.26
LARCM + Added Value	81.06 ± 6.13	82.70 ± 5.42	71.64 ± 5.79	83.66 ± 5.07	84.19 ± 5.24	76.80 ± 6.39	77.84 ± 5.91	77.78 ± 6.00
LARCM + Certainty Factor	83.31 ± 5.89	84.47 ± 5.77	75.41 ± 5.95	86.48 ± 5.56	86.11 ± 4.91	77.20 ± 6.00	78.21 ± 6.07	79.95 ± 6.22
LARCM + Collective Strength	80.87 ± 5.52	83.43 ± 5.17	73.51 ± 6.40	84.34 ± 5.17	86.35 ± 4.97	75.82 ± 5.81	76.76 ± 5.91	77.65 ± 5.85
LARCM + Confidence	80.69 ± 5.24	83.11 ± 4.96	73.79 ± 6.49	84.83 ± 4.46	85.03 ± 4.66	76.42 ± 5.70	78.17 ± 4.87	79.16 ± 4.90
LARCM + Conviction	68.14 ± 6.94	75.73 ± 6.70	65.41 ± 7.89	74.64 ± 6.98	74.16 ± 6.80	70.27 ± 7.08	70.15 ± 7.13	69.88 ± 6.85
LARCM + ϕ-Coefficient	82.34 ± 5.40	85.38 ± 4.83	74.61 ± 7.03	86.53 ± 5.14	86.32 ± 4.55	81.05 ± 5.50	82.78 ± 5.71	82.38 ± 5.61
LARCM + Gini Index	82.22 ± 5.68	85.75 ± 5.07	74.78 ± 6.69	87.29 ± 4.59	86.92 ± 4.45	77.06 ± 5.66	79.00 ± 4.97	78.66 ± 5.63
LARCM + IS	82.33 ± 5.49	83.61 ± 5.86	75.26 ± 5.44	85.22 ± 5.47	85.46 ± 5.52	76.00 ± 6.20	77.47 ± 6.26	78.78 ± 6.25
LARCM + J-Measure	81.43 ± 5.37	84.31 ± 5.14	73.83 ± 6.48	85.17 ± 4.56	84.45 ± 5.28	79.55 ± 5.85	80.28 ± 5.61	80.52 ± 5.96
LARCM + Kappa	80.02 ± 5.93	82.26 ± 4.69	76.09 ± 6.28	84.23 ± 5.09	85.56 ± 4.86	76.54 ± 6.22	77.06 ± 5.59	77.72 ± 5.62
LARCM + Klosgen	82.31 ± 5.53	85.07 ± 5.59	73.23 ± 6.70	86.58 ± 5.13	86.59 ± 4.52	80.08 ± 5.80	80.38 ± 5.20	81.06 ± 5.27
LARCM + Lambda	81.52 ± 5.58	83.66 ± 5.42	72.26 ± 7.07	85.39 ± 4.58	83.47 ± 5.06	74.52 ± 5.57	76.51 ± 6.04	76.72 ± 5.74
LARCM + Laplace	83.57 ± 5.53	86.41 ± 4.71	74.87 ± 6.62	85.89 ± 5.12	86.81 ± 4.97	79.13 ± 5.65	80.05 ± 5.32	80.84 ± 5.18
LARCM + Lift	80.75 ± 5.86	84.34 ± 5.86	75.60 ± 7.41	86.06 ± 5.41	86.44 ± 5.11	75.67 ± 6.32	78.67 ± 6.54	79.88 ± 6.38
LARCM + Mutual Information LHS	82.76 ± 5.92	85.49 ± 4.82	74.21 ± 5.87	85.54 ± 5.13	85.90 ± 5.81	77.83 ± 5.72	78.67 ± 5.21	79.52 ± 5.39
LARCM + Novelty	83.57 ± 6.30	84.33 ± 5.46	75.14 ± 6.73	85.26 ± 5.49	87.30 ± 5.06	79.06 ± 6.34	79.76 ± 6.93	80.60 ± 6.57
LARCM + Support	86.75 ± 4.73	85.36 ± 5.38	80.76 ± 5.72	87.65 ± 4.81	90.34 ± 4.43	86.06 ± 5.78	86.90 ± 5.46	85.89 ± 5.41

Table A8
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-3.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	84.19 ± 4.62	82.29 ± 4.79	80.73 ± 5.99	81.66 ± 4.92	84.15 ± 4.51	70.36 ± 5.30	72.04 ± 5.96	74.18 ± 6.06
LDA + *bag-of-words*	84.13 ± 5.23	87.17 ± 5.18	77.40 ± 5.72	85.00 ± 5.17	84.96 ± 5.03	82.11 ± 5.25	82.99 ± 5.18	81.20 ± 5.36
LARCM + Added Value	83.05 ± 6.19	82.50 ± 5.64	80.31 ± 5.72	86.98 ± 5.16	87.55 ± 4.98	79.19 ± 6.71	78.90 ± 6.70	80.12 ± 6.75
LARCM + Certainty Factor	84.71 ± 5.81	83.81 ± 5.37	74.32 ± 6.31	87.05 ± 4.71	85.60 ± 5.32	77.29 ± 6.05	79.84 ± 5.69	80.33 ± 5.97
LARCM + Collective Strength	85.45 ± 5.04	84.40 ± 4.65	76.80 ± 6.29	87.14 ± 4.60	87.61 ± 4.17	78.74 ± 6.15	80.65 ± 5.11	82.06 ± 4.79
LARCM + Confidence	84.24 ± 5.42	82.72 ± 6.00	79.30 ± 5.66	85.03 ± 5.04	85.32 ± 5.04	76.76 ± 6.39	78.18 ± 6.45	78.28 ± 6.29
LARCM + Conviction	66.17 ± 6.64	73.51 ± 7.37	63.61 ± 6.69	71.79 ± 7.11	71.81 ± 7.49	68.65 ± 7.87	70.47 ± 7.51	70.64 ± 7.16
LARCM + ϕ-Coefficient	84.45 ± 4.82	85.37 ± 4.79	79.47 ± 6.33	87.43 ± 4.92	87.65 ± 4.67	80.05 ± 6.24	81.40 ± 6.04	81.74 ± 6.28
LARCM + Gini Index	84.38 ± 5.28	82.59 ± 5.48	77.64 ± 5.56	87.08 ± 4.60	86.84 ± 5.01	77.68 ± 6.07	79.28 ± 5.81	79.18 ± 5.93
LARCM + IS	82.48 ± 5.23	81.86 ± 5.65	77.39 ± 5.79	85.44 ± 4.67	83.42 ± 4.81	75.26 ± 6.01	77.66 ± 5.54	78.52 ± 5.52
LARCM + J-Measure	84.16 ± 4.53	81.88 ± 5.36	78.16 ± 6.26	86.01 ± 4.74	86.51 ± 4.18	74.92 ± 6.46	76.60 ± 6.11	79.15 ± 5.60
LARCM + Kappa	86.87 ± 4.97	82.53 ± 5.35	76.76 ± 6.67	85.61 ± 5.07	86.69 ± 4.94	77.27 ± 6.45	78.54 ± 5.71	77.77 ± 5.59
LARCM + Klosgen	84.43 ± 5.41	83.53 ± 5.28	75.05 ± 5.77	85.84 ± 5.05	84.88 ± 5.15	75.68 ± 6.06	79.67 ± 5.46	79.54 ± 6.34
LARCM + Lambda	82.09 ± 5.10	82.76 ± 5.21	74.69 ± 6.51	85.94 ± 5.03	85.05 ± 5.55	72.57 ± 6.45	74.92 ± 6.36	76.69 ± 6.44
LARCM + Laplace	84.62 ± 4.87	82.67 ± 5.46	76.13 ± 6.12	86.41 ± 4.95	86.97 ± 5.48	75.34 ± 6.45	78.42 ± 6.53	78.99 ± 6.46
LARCM + Lift	83.85 ± 5.28	81.11 ± 5.04	75.41 ± 6.61	86.18 ± 4.94	85.72 ± 4.55	73.39 ± 6.03	76.25 ± 6.28	76.37 ± 6.11
LARCM + Mutual Information LHS	82.52 ± 5.50	83.00 ± 5.17	76.07 ± 5.77	84.88 ± 5.29	84.92 ± 5.80	73.57 ± 5.37	74.87 ± 6.03	76.43 ± 5.72
LARCM + Novelty	86.47 ± 4.72	84.16 ± 5.12	75.92 ± 6.74	86.08 ± 4.91	87.38 ± 5.32	78.25 ± 5.47	78.42 ± 5.41	80.03 ± 5.32
LARCM + Support	88.99 ± 4.25	86.09 ± 5.07	82.15 ± 5.06	89.19 ± 4.63	91.11 ± 4.43	83.23 ± 5.02	86.07 ± 4.97	86.33 ± 4.73

Table A9
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-3.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	76.95 ± 6.25	78.36 ± 6.64	67.85 ± 6.46	78.27 ± 6.48	81.06 ± 6.18	66.61 ± 7.89	66.42 ± 8.31	58.70 ± 7.90
LDA + *bag-of-words*	77.33 ± 5.61	80.68 ± 5.56	72.79 ± 5.42	65.29 ± 6.53	81.32 ± 5.78	71.02 ± 6.92	76.68 ± 5.22	73.98 ± 6.21
LARCM + Added Value	83.73 ± 5.56	80.51 ± 6.12	80.27 ± 6.44	83.92 ± 5.07	85.41 ± 5.38	74.21 ± 6.70	74.95 ± 6.48	76.40 ± 6.09
LARCM + Certainty Factor	84.92 ± 5.90	81.08 ± 5.96	79.33 ± 5.67	83.94 ± 5.88	86.27 ± 5.96	75.48 ± 6.25	76.25 ± 6.62	76.89 ± 6.00
LARCM + Collective Strength	83.13 ± 5.72	81.42 ± 5.87	75.85 ± 5.59	85.49 ± 5.46	85.72 ± 5.79	75.72 ± 5.26	77.15 ± 5.39	77.24 ± 5.18
LARCM + Confidence	84.86 ± 5.62	81.81 ± 5.37	81.05 ± 7.08	84.53 ± 5.56	85.73 ± 5.36	77.22 ± 5.85	79.14 ± 5.58	79.16 ± 5.73
LARCM + Conviction	65.10 ± 6.81	72.89 ± 6.07	62.00 ± 7.29	72.91 ± 5.78	72.43 ± 5.93	62.55 ± 7.58	61.42 ± 7.74	59.50 ± 7.50
LARCM + ϕ-Coefficient	86.84 ± 5.19	82.30 ± 6.41	79.70 ± 6.35	87.45 ± 5.43	87.17 ± 4.77	80.96 ± 6.08	81.89 ± 6.32	82.63 ± 5.94
LARCM + Gini Index	84.68 ± 5.99	80.35 ± 6.20	79.22 ± 6.96	85.57 ± 5.88	86.87 ± 5.19	74.60 ± 6.11	76.86 ± 6.66	77.34 ± 6.96
LARCM + IS	84.67 ± 5.42	79.69 ± 5.61	78.44 ± 6.29	84.98 ± 4.99	85.07 ± 4.71	75.94 ± 6.78	79.16 ± 5.52	80.05 ± 5.70
LARCM + J-Measure	83.58 ± 5.45	81.55 ± 5.71	78.40 ± 6.11	84.60 ± 5.79	86.33 ± 4.94	75.33 ± 6.45	78.37 ± 5.55	79.32 ± 5.80
LARCM + Kappa	83.67 ± 4.93	81.95 ± 5.37	77.69 ± 6.55	86.08 ± 5.31	87.16 ± 4.72	78.16 ± 6.19	80.14 ± 5.52	80.06 ± 5.48
LARCM + Klosgen	83.90 ± 5.27	80.82 ± 5.37	76.51 ± 5.71	85.03 ± 5.40	86.81 ± 5.03	76.25 ± 6.23	76.98 ± 6.55	78.93 ± 6.29
LARCM + Lambda	82.08 ± 5.53	81.45 ± 5.72	72.50 ± 7.01	85.78 ± 5.61	85.43 ± 5.58	72.85 ± 6.60	72.87 ± 6.87	75.74 ± 7.41
LARCM + Laplace	87.09 ± 4.89	82.44 ± 4.88	79.38 ± 6.20	84.12 ± 5.31	86.25 ± 4.89	74.93 ± 6.02	78.40 ± 5.74	78.21 ± 5.71
LARCM + Lift	84.14 ± 4.65	80.77 ± 5.25	76.36 ± 6.86	85.86 ± 4.81	86.51 ± 4.85	76.54 ± 6.54	78.53 ± 6.05	79.21 ± 5.91
LARCM + Mutual Information LHS	85.76 ± 5.22	78.48 ± 6.33	75.91 ± 5.98	85.01 ± 6.05	85.07 ± 5.83	75.74 ± 7.29	76.73 ± 6.97	78.06 ± 6.92
LARCM + Novelty	86.91 ± 4.71	81.91 ± 6.05	80.07 ± 5.54	84.66 ± 5.59	86.48 ± 5.21	78.11 ± 6.18	79.07 ± 5.69	77.60 ± 5.81
LARCM + Support	91.11 ± 4.26	83.81 ± 5.69	85.69 ± 5.48	88.02 ± 4.90	92.64 ± 4.09	84.03 ± 6.44	85.96 ± 5.51	85.57 ± 5.36

Table A10
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-4.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	93.56 ± 3.57	91.59 ± 4.13	91.78 ± 4.56	90.78 ± 4.27	93.68 ± 3.98	83.82 ± 5.12	84.84 ± 4.98	85.63 ± 5.30
LDA + *bag-of-words*	94.85 ± 3.69	91.87 ± 3.79	92.15 ± 4.40	95.58 ± 3.13	94.82 ± 3.16	92.41 ± 4.00	92.26 ± 4.14	93.45 ± 3.81
LARCM + Added Value	89.19 ± 4.15	84.61 ± 4.12	85.20 ± 5.20	88.22 ± 3.96	91.80 ± 4.13	84.95 ± 4.70	86.42 ± 4.58	87.14 ± 4.72
LARCM + Certainty Factor	87.88 ± 4.29	85.19 ± 3.88	82.58 ± 6.06	88.33 ± 4.81	91.28 ± 3.93	83.91 ± 5.34	86.60 ± 5.10	87.39 ± 4.74
LARCM + Collective Strength	89.87 ± 4.15	85.48 ± 3.59	84.41 ± 5.78	89.36 ± 4.20	92.56 ± 3.41	87.69 ± 4.36	88.14 ± 4.23	88.37 ± 4.46
LARCM + Confidence	88.91 ± 4.93	85.11 ± 4.30	87.67 ± 4.79	88.56 ± 4.86	90.79 ± 4.56	85.19 ± 6.13	86.79 ± 5.90	86.89 ± 5.94
LARCM + Conviction	76.48 ± 6.97	79.87 ± 5.68	73.81 ± 6.68	82.57 ± 6.38	83.12 ± 5.80	78.25 ± 6.56	79.95 ± 6.25	80.39 ± 6.10
LARCM + ϕ-Coefficient	88.17 ± 4.06	86.57 ± 3.84	81.80 ± 4.83	90.61 ± 4.19	91.82 ± 4.28	87.20 ± 4.78	88.07 ± 4.94	88.32 ± 4.78
LARCM + Gini Index	86.13 ± 5.11	85.91 ± 3.93	83.31 ± 4.80	89.54 ± 4.22	91.69 ± 4.24	86.21 ± 5.25	86.59 ± 4.92	87.12 ± 5.05
LARCM + IS	87.52 ± 5.70	83.83 ± 4.61	83.75 ± 5.54	87.61 ± 4.84	90.69 ± 4.34	85.10 ± 5.23	85.43 ± 5.38	85.10 ± 5.46
LARCM + J-Measure	88.17 ± 4.30	86.40 ± 3.94	83.76 ± 5.16	89.87 ± 4.16	92.11 ± 3.82	86.44 ± 4.82	87.00 ± 4.84	87.05 ± 4.93
LARCM + Kappa	88.35 ± 4.82	84.03 ± 4.56	85.46 ± 5.61	86.62 ± 4.65	90.77 ± 4.63	86.96 ± 4.95	85.48 ± 4.83	86.17 ± 4.83
LARCM + Klosgen	85.26 ± 5.33	85.99 ± 3.96	85.32 ± 4.99	88.99 ± 4.54	90.79 ± 4.10	84.29 ± 5.58	85.51 ± 5.90	85.50 ± 5.65
LARCM + Lambda	83.95 ± 4.95	84.55 ± 3.92	80.80 ± 5.05	87.74 ± 4.13	89.27 ± 4.40	85.91 ± 4.09	86.30 ± 4.59	86.25 ± 4.81
LARCM + Laplace	86.65 ± 4.78	84.34 ± 4.24	86.14 ± 4.59	88.43 ± 4.54	91.63 ± 3.90	86.40 ± 5.44	86.87 ± 5.19	86.73 ± 5.57
LARCM + Lift	88.28 ± 4.25	84.11 ± 4.60	82.31 ± 4.94	88.74 ± 4.46	91.55 ± 3.84	85.76 ± 4.76	86.54 ± 4.86	86.98 ± 4.76
LARCM + Mutual Information LHS	88.45 ± 4.52	84.71 ± 4.49	86.25 ± 5.39	88.22 ± 5.00	89.32 ± 4.43	86.57 ± 5.06	86.52 ± 4.71	86.14 ± 5.23
LARCM + Novelty	89.97 ± 4.74	85.81 ± 4.62	83.48 ± 6.05	89.44 ± 4.73	91.53 ± 4.12	85.61 ± 5.31	87.11 ± 5.27	87.31 ± 5.11
LARCM + Support	92.18 ± 4.39	89.62 ± 3.83	89.60 ± 4.30	94.26 ± 3.32	94.74 ± 3.19	92.23 ± 3.76	92.74 ± 3.69	92.96 ± 3.72

Table A11
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-4.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	89.52 ± 4.78	86.75 ± 5.18	89.96 ± 4.71	89.12 ± 4.99	93.28 ± 3.73	80.79 ± 6.02	82.95 ± 5.97	83.76 ± 5.49
LDA + *bag-of-words*	90.68 ± 4.39	81.92 ± 3.94	83.66 ± 5.36	87.78 ± 3.99	91.40 ± 3.76	86.52 ± 4.53	87.73 ± 5.25	86.46 ± 5.62
LARCM + Added Value	88.35 ± 4.67	82.28 ± 3.86	85.36 ± 5.40	84.21 ± 4.63	90.10 ± 3.86	82.07 ± 5.79	84.81 ± 5.68	84.99 ± 5.11
LARCM + Certainty Factor	90.30 ± 4.51	83.17 ± 4.52	87.68 ± 5.25	86.30 ± 4.98	89.83 ± 4.33	83.55 ± 5.50	84.87 ± 5.03	86.07 ± 4.99
LARCM + Collective Strength	88.47 ± 4.95	83.00 ± 3.62	85.00 ± 5.64	87.14 ± 4.49	90.03 ± 3.98	86.22 ± 5.29	85.43 ± 5.43	85.73 ± 5.44
LARCM + Confidence	90.36 ± 4.12	82.79 ± 4.50	90.02 ± 4.32	88.51 ± 4.23	91.12 ± 4.20	84.22 ± 5.24	85.18 ± 5.62	85.57 ± 5.25
LARCM + Conviction	73.41 ± 5.86	78.07 ± 5.34	73.98 ± 7.05	77.38 ± 5.81	80.76 ± 6.41	74.94 ± 6.28	78.83 ± 5.91	79.71 ± 5.79
LARCM + ϕ-Coefficient	91.83 ± 4.05	84.67 ± 2.83	87.73 ± 5.27	87.88 ± 4.60	91.83 ± 4.14	86.07 ± 4.79	87.64 ± 5.04	86.75 ± 4.64
LARCM + Gini Index	91.02 ± 4.36	83.72 ± 3.95	85.30 ± 4.91	87.41 ± 4.27	91.09 ± 4.18	85.71 ± 5.08	85.66 ± 4.46	85.66 ± 5.01
LARCM + IS	90.73 ± 4.17	84.06 ± 3.58	86.04 ± 4.67	86.35 ± 4.60	90.46 ± 4.52	84.09 ± 5.30	83.83 ± 5.06	84.06 ± 4.99
LARCM + J-Measure	89.03 ± 4.35	83.75 ± 3.90	87.24 ± 5.25	87.67 ± 4.36	90.58 ± 4.04	82.92 ± 5.27	82.13 ± 5.04	83.45 ± 4.79
LARCM + Kappa	90.09 ± 4.72	83.75 ± 4.03	85.85 ± 5.20	86.07 ± 4.60	90.89 ± 5.01	81.45 ± 5.07	82.08 ± 4.57	82.79 ± 5.34
LARCM + Klosgen	90.12 ± 4.42	82.72 ± 3.32	89.00 ± 4.65	85.05 ± 4.21	90.88 ± 4.32	82.52 ± 5.45	85.40 ± 5.40	86.65 ± 5.04
LARCM + Lambda	86.62 ± 5.17	82.79 ± 4.34	84.66 ± 5.83	85.62 ± 5.25	90.88 ± 5.18	82.92 ± 5.98	85.59 ± 5.24	85.55 ± 5.22
LARCM + Laplace	89.90 ± 3.84	83.10 ± 4.09	87.82 ± 4.23	87.60 ± 4.77	90.77 ± 4.49	85.74 ± 5.16	86.45 ± 5.25	86.80 ± 5.34
LARCM + Lift	89.54 ± 4.38	82.26 ± 4.28	86.81 ± 5.02	87.79 ± 4.38	90.94 ± 4.18	82.38 ± 5.31	82.36 ± 5.45	83.32 ± 5.33
LARCM + Mutual Information LHS	89.74 ± 4.16	84.29 ± 3.50	86.77 ± 5.66	88.30 ± 4.35	91.09 ± 4.70	82.21 ± 5.88	83.27 ± 5.81	84.26 ± 6.20
LARCM + Novelty	90.63 ± 4.54	83.34 ± 3.72	88.96 ± 4.35	84.89 ± 4.36	91.39 ± 3.55	85.57 ± 4.69	86.15 ± 4.82	86.26 ± 4.82
LARCM + Support	92.76 ± 4.23	85.57 ± 3.77	90.02 ± 5.04	91.08 ± 4.18	94.49 ± 3.39	92.22 ± 4.44	91.28 ± 4.49	90.64 ± 4.30

Table A12
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-4.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	85.26 ± 5.60	79.36 ± 4.93	80.09 ± 5.71	85.07 ± 5.35	88.02 ± 4.83	76.61 ± 6.54	70.73 ± 6.09	66.22 ± 6.56
LDA + *bag-of-words*	89.23 ± 4.53	73.68 ± 5.09	80.72 ± 6.36	81.51 ± 5.46	88.23 ± 5.07	85.60 ± 5.47	84.10 ± 5.68	82.68 ± 5.62
LARCM + Added Value	91.73 ± 4.29	79.81 ± 4.14	88.47 ± 4.86	81.97 ± 4.49	90.55 ± 4.71	86.52 ± 5.44	87.65 ± 5.53	87.71 ± 5.10
LARCM + Certainty Factor	89.90 ± 5.01	81.98 ± 4.15	84.52 ± 6.20	85.89 ± 4.92	91.38 ± 3.87	86.14 ± 5.18	86.09 ± 5.59	87.10 ± 4.69
LARCM + Collective Strength	92.06 ± 4.35	81.34 ± 3.64	85.16 ± 4.75	81.98 ± 4.16	92.54 ± 3.57	86.27 ± 4.89	87.11 ± 5.05	86.71 ± 5.25
LARCM + Confidence	90.38 ± 3.80	80.92 ± 3.78	87.94 ± 4.88	84.50 ± 4.79	90.71 ± 4.44	83.73 ± 4.67	84.92 ± 4.59	84.75 ± 4.65
LARCM + Conviction	71.14 ± 6.78	76.73 ± 4.68	72.01 ± 6.84	76.80 ± 5.20	80.57 ± 6.29	74.96 ± 6.35	73.52 ± 6.82	71.44 ± 7.65
LARCM + ϕ-Coefficient	91.27 ± 4.55	81.93 ± 3.77	87.64 ± 4.63	86.14 ± 4.61	91.01 ± 4.08	84.04 ± 5.82	84.41 ± 5.05	84.61 ± 4.77
LARCM + Gini Index	92.17 ± 4.19	80.96 ± 4.01	86.85 ± 4.99	81.85 ± 4.60	90.26 ± 4.19	82.60 ± 5.69	81.86 ± 5.56	82.55 ± 5.51
LARCM + IS	91.50 ± 4.42	81.95 ± 3.32	83.85 ± 5.23	83.32 ± 4.70	91.30 ± 4.74	82.88 ± 6.41	84.00 ± 5.23	85.33 ± 4.83
LARCM + J-Measure	90.51 ± 4.00	80.86 ± 3.53	86.60 ± 5.03	81.55 ± 3.85	91.25 ± 4.13	78.24 ± 5.61	80.51 ± 5.18	82.33 ± 5.11
LARCM + Kappa	90.53 ± 4.29	81.34 ± 3.59	87.82 ± 4.79	81.37 ± 3.97	90.23 ± 4.26	84.24 ± 4.98	84.38 ± 5.24	84.56 ± 4.73
LARCM + Klosgen	91.14 ± 4.07	79.98 ± 3.68	86.52 ± 5.00	82.59 ± 3.63	90.71 ± 4.33	85.89 ± 5.32	86.39 ± 5.22	86.90 ± 5.21
LARCM + Lambda	88.88 ± 4.57	81.07 ± 4.08	83.51 ± 5.50	82.22 ± 5.08	88.13 ± 4.63	80.30 ± 5.91	82.87 ± 5.36	83.81 ± 5.16
LARCM + Laplace	91.75 ± 4.42	81.32 ± 3.69	85.66 ± 4.92	83.30 ± 4.50	90.18 ± 4.45	83.86 ± 5.19	83.49 ± 5.19	82.81 ± 5.45
LARCM + Lift	91.06 ± 4.18	79.92 ± 4.16	82.57 ± 5.52	81.16 ± 4.14	89.36 ± 4.86	79.46 ± 5.30	81.95 ± 5.29	83.19 ± 4.65
LARCM + Mutual Information LHS	89.67 ± 4.16	81.71 ± 3.67	86.17 ± 5.18	83.33 ± 4.11	90.43 ± 4.42	83.99 ± 4.96	84.57 ± 5.69	85.20 ± 5.24
LARCM + Novelty	89.34 ± 4.59	81.01 ± 3.69	87.66 ± 4.93	81.79 ± 3.94	89.34 ± 3.80	82.18 ± 5.54	84.01 ± 5.12	83.73 ± 5.34
LARCM + Support	94.26 ± 3.29	83.57 ± 3.53	91.07 ± 4.39	88.39 ± 4.32	94.49 ± 3.40	91.85 ± 4.39	91.93 ± 4.46	91.30 ± 4.64

Table A13
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-5.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	81.67 ± 5.21	85.80 ± 5.29	85.22 ± 4.80	84.14 ± 6.02	87.33 ± 4.95	78.61 ± 5.85	77.19 ± 6.30	77.21 ± 6.41
LDA + *bag-of-words*	86.62 ± 4.64	90.13 ± 4.33	88.09 ± 4.28	89.72 ± 4.15	90.59 ± 3.94	83.31 ± 4.79	83.86 ± 5.05	84.66 ± 4.52
LARCM + Added Value	72.18 ± 6.61	76.45 ± 4.80	69.62 ± 5.06	77.25 ± 5.08	77.89 ± 5.17	67.91 ± 6.60	69.85 ± 6.26	71.65 ± 5.81
LARCM + Certainty Factor	73.80 ± 5.43	76.48 ± 5.81	68.75 ± 5.74	78.62 ± 5.41	78.15 ± 4.91	71.36 ± 5.77	73.21 ± 5.48	73.74 ± 5.18
LARCM + Collective Strength	72.44 ± 5.25	76.75 ± 5.64	66.99 ± 6.37	77.64 ± 5.84	77.56 ± 5.63	70.78 ± 5.79	73.62 ± 5.89	73.43 ± 5.43
LARCM + Confidence	74.33 ± 5.58	77.22 ± 4.72	71.09 ± 6.35	79.39 ± 4.96	78.98 ± 4.93	72.91 ± 6.21	73.74 ± 5.97	73.89 ± 6.21
LARCM + Conviction	62.38 ± 6.14	71.66 ± 6.01	60.44 ± 6.69	70.83 ± 6.10	70.34 ± 6.74	65.89 ± 5.68	66.71 ± 5.49	68.26 ± 5.80
LARCM + ϕ-Coefficient	74.03 ± 5.93	79.08 ± 4.98	70.17 ± 6.00	79.89 ± 5.53	80.59 ± 6.13	73.25 ± 5.58	75.07 ± 5.77	75.58 ± 6.01
LARCM + Gini Index	72.23 ± 6.15	77.41 ± 5.50	68.94 ± 6.28	79.36 ± 5.82	78.68 ± 5.62	73.23 ± 6.91	74.60 ± 6.47	75.39 ± 6.64
LARCM + IS	72.46 ± 5.83	77.84 ± 5.74	66.57 ± 6.56	77.86 ± 5.51	77.97 ± 5.49	71.02 ± 6.02	73.12 ± 5.15	73.70 ± 5.60
LARCM + J-Measure	70.65 ± 5.53	76.87 ± 5.19	68.83 ± 6.41	77.62 ± 5.37	77.19 ± 5.44	72.25 ± 6.09	73.37 ± 6.38	73.80 ± 5.58
LARCM + Kappa	70.65 ± 6.51	76.35 ± 5.73	67.17 ± 6.72	77.75 ± 5.84	76.07 ± 6.17	71.51 ± 6.11	73.72 ± 5.50	73.12 ± 5.79
LARCM + Klosgen	72.72 ± 5.18	76.73 ± 5.05	67.83 ± 6.87	77.64 ± 5.54	77.90 ± 6.36	69.79 ± 6.44	71.25 ± 5.89	71.82 ± 6.10
LARCM + Lambda	72.66 ± 6.25	78.22 ± 5.50	68.66 ± 6.59	78.54 ± 5.62	78.20 ± 5.66	71.47 ± 6.10	72.77 ± 6.23	74.10 ± 6.51
LARCM + Laplace	73.13 ± 6.23	77.08 ± 4.92	70.24 ± 5.86	77.71 ± 4.88	76.65 ± 5.74	70.72 ± 5.51	72.21 ± 5.47	74.38 ± 5.36
LARCM + Lift	73.72 ± 5.38	78.20 ± 4.99	67.84 ± 5.29	78.63 ± 4.90	79.13 ± 5.39	75.33 ± 5.59	74.95 ± 5.36	76.01 ± 5.19
LARCM + Mutual Information LHS	72.06 ± 6.08	76.90 ± 5.38	67.62 ± 5.64	76.80 ± 5.86	78.11 ± 5.65	72.13 ± 6.16	72.27 ± 6.14	73.50 ± 6.36
LARCM + Novelty	72.93 ± 5.73	76.82 ± 5.32	72.99 ± 6.30	77.78 ± 5.28	78.27 ± 5.58	73.78 ± 5.62	74.02 ± 5.68	74.36 ± 6.08
LARCM + Support	78.88 ± 4.70	81.60 ± 4.79	74.00 ± 5.81	82.19 ± 4.35	84.44 ± 4.75	78.75 ± 5.29	80.28 ± 4.28	81.22 ± 4.40

Table A14
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-5.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	80.68 ± 5.28	77.81 ± 4.97	80.59 ± 5.78	79.94 ± 5.53	82.18 ± 5.58	64.56 ± 6.07	62.06 ± 6.64	58.53 ± 5.93
LDA + *bag-of-words*	86.56 ± 4.85	85.81 ± 4.56	85.90 ± 4.94	86.43 ± 4.84	88.68 ± 4.58	79.15 ± 5.87	79.85 ± 5.47	79.24 ± 5.18
LARCM + Added Value	75.54 ± 5.93	78.51 ± 4.71	71.45 ± 5.94	76.88 ± 5.06	78.21 ± 5.51	71.09 ± 5.91	72.65 ± 5.65	73.27 ± 6.15
LARCM + Certainty Factor	76.90 ± 4.90	78.71 ± 4.84	71.96 ± 6.25	78.89 ± 5.53	80.06 ± 5.20	73.97 ± 6.01	75.12 ± 6.39	76.56 ± 5.58
LARCM + Collective Strength	73.90 ± 6.59	77.43 ± 4.95	68.58 ± 6.50	77.98 ± 5.35	77.92 ± 6.02	70.17 ± 6.00	73.59 ± 6.18	73.89 ± 6.87
LARCM + Confidence	74.72 ± 6.47	78.18 ± 5.33	71.66 ± 5.82	78.15 ± 5.28	79.30 ± 5.47	71.78 ± 5.67	72.14 ± 5.82	71.91 ± 5.85
LARCM + Conviction	59.51 ± 6.18	70.87 ± 5.67	60.78 ± 6.31	70.74 ± 6.66	71.08 ± 6.28	64.97 ± 6.63	65.90 ± 6.87	67.62 ± 6.21
LARCM + ϕ-Coefficient	77.66 ± 6.11	77.35 ± 5.23	71.05 ± 5.40	78.28 ± 5.11	79.32 ± 5.57	71.57 ± 6.09	74.01 ± 5.30	75.16 ± 5.41
LARCM + Gini Index	74.37 ± 5.35	77.63 ± 5.00	70.68 ± 6.05	77.35 ± 6.02	78.96 ± 5.66	74.26 ± 5.35	73.52 ± 5.10	74.52 ± 5.54
LARCM + IS	72.06 ± 5.48	77.33 ± 4.89	69.39 ± 5.98	76.94 ± 4.95	77.90 ± 5.55	70.38 ± 6.32	71.15 ± 5.87	72.70 ± 6.53
LARCM + J-Measure	74.37 ± 5.61	78.32 ± 4.96	70.91 ± 6.95	77.26 ± 5.46	78.43 ± 5.23	70.48 ± 6.17	72.56 ± 6.12	72.80 ± 5.49
LARCM + Kappa	75.20 ± 5.32	78.12 ± 5.28	68.83 ± 6.19	77.73 ± 5.55	77.65 ± 6.14	68.30 ± 5.88	70.47 ± 5.31	71.41 ± 5.72
LARCM + Klosgen	76.15 ± 5.72	77.78 ± 5.24	72.82 ± 5.64	77.48 ± 5.10	77.22 ± 5.13	70.55 ± 6.21	72.99 ± 5.54	73.61 ± 5.57
LARCM + Lambda	73.01 ± 6.76	78.98 ± 4.97	67.81 ± 6.83	78.79 ± 5.04	79.43 ± 5.84	69.51 ± 6.78	70.35 ± 6.78	70.34 ± 6.60
LARCM + Laplace	76.20 ± 6.03	77.85 ± 5.02	72.76 ± 6.04	78.02 ± 5.37	78.04 ± 5.80	70.40 ± 6.25	71.21 ± 6.57	73.65 ± 5.69
LARCM + Lift	73.57 ± 6.27	77.50 ± 4.97	67.60 ± 6.91	79.07 ± 5.23	79.26 ± 5.30	71.07 ± 6.24	72.82 ± 5.64	73.23 ± 5.28
LARCM + Mutual Information LHS	74.78 ± 5.83	77.62 ± 5.00	70.15 ± 6.14	77.32 ± 5.40	78.60 ± 5.30	70.36 ± 6.11	72.39 ± 5.78	73.95 ± 6.03
LARCM + Novelty	74.82 ± 5.48	78.28 ± 5.42	70.90 ± 5.74	78.19 ± 5.60	79.81 ± 5.43	71.32 ± 6.47	72.51 ± 6.29	73.82 ± 5.98
LARCM + Support	83.38 ± 5.10	80.49 ± 5.30	77.73 ± 5.41	82.06 ± 5.39	84.82 ± 5.10	78.84 ± 5.68	80.91 ± 5.21	81.79 ± 5.68

Table A15
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-5.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	76.45 ± 5.71	77.37 ± 5.91	78.87 ± 5.95	76.44 ± 5.23	80.75 ± 6.06	57.47 ± 6.95	58.78 ± 7.79	62.42 ± 6.02
LDA + *bag-of-words*	84.25 ± 4.72	80.90 ± 4.72	76.75 ± 5.53	71.30 ± 5.81	82.66 ± 5.08	67.83 ± 5.50	70.15 ± 6.30	67.49 ± 6.44
LARCM + Added Value	77.43 ± 5.50	78.01 ± 5.43	74.75 ± 6.04	76.82 ± 5.76	80.43 ± 6.03	70.73 ± 5.70	72.57 ± 5.83	73.33 ± 6.19
LARCM + Certainty Factor	75.99 ± 5.70	78.70 ± 5.01	71.52 ± 5.99	76.56 ± 5.07	79.73 ± 5.77	70.78 ± 6.52	71.93 ± 6.66	73.20 ± 7.17
LARCM + Collective Strength	77.41 ± 5.35	77.61 ± 4.97	68.79 ± 5.88	77.94 ± 5.20	79.54 ± 5.09	72.66 ± 6.13	73.78 ± 6.25	75.14 ± 6.03
LARCM + Confidence	80.34 ± 5.02	77.71 ± 4.69	73.17 ± 6.60	77.18 ± 4.53	80.09 ± 4.97	69.28 ± 5.53	70.38 ± 4.90	72.23 ± 5.38
LARCM + Conviction	64.36 ± 5.33	70.43 ± 5.99	58.59 ± 6.54	70.26 ± 6.09	69.07 ± 6.33	62.76 ± 6.29	63.40 ± 5.54	63.38 ± 6.38
LARCM + ϕ-Coefficient	79.79 ± 4.92	79.04 ± 4.48	71.89 ± 6.22	78.85 ± 4.78	79.79 ± 5.34	71.21 ± 6.65	73.35 ± 6.63	75.01 ± 5.98
LARCM + Gini Index	76.47 ± 5.64	77.98 ± 5.47	74.55 ± 6.04	76.56 ± 5.94	80.21 ± 5.83	70.17 ± 6.50	71.48 ± 7.25	73.25 ± 6.47
LARCM + IS	76.33 ± 5.51	77.39 ± 5.22	69.81 ± 6.05	75.37 ± 5.80	78.55 ± 5.29	68.42 ± 6.43	70.97 ± 5.97	72.27 ± 5.40
LARCM + J-Measure	76.37 ± 5.52	77.69 ± 4.18	70.28 ± 6.28	77.35 ± 4.61	80.15 ± 5.65	69.79 ± 6.31	70.93 ± 6.90	72.30 ± 6.56
LARCM + Kappa	79.74 ± 5.35	77.43 ± 5.14	72.21 ± 6.89	76.75 ± 4.98	78.19 ± 5.91	69.96 ± 5.99	72.88 ± 6.20	72.84 ± 5.98
LARCM + Klosgen	75.78 ± 5.65	77.71 ± 4.81	69.72 ± 6.41	75.88 ± 5.24	79.54 ± 5.92	70.92 ± 5.79	71.83 ± 6.60	72.59 ± 6.65
LARCM + Lambda	74.86 ± 4.84	77.77 ± 5.21	67.06 ± 5.72	76.11 ± 5.03	79.55 ± 5.76	66.24 ± 5.50	67.32 ± 6.08	69.89 ± 5.34
LARCM + Laplace	80.51 ± 5.04	78.50 ± 4.99	72.38 ± 5.83	77.12 ± 5.08	80.85 ± 5.23	67.98 ± 6.42	69.19 ± 6.45	71.57 ± 6.22
LARCM + Lift	78.03 ± 6.13	77.50 ± 5.30	66.47 ± 7.35	76.27 ± 5.43	78.60 ± 6.41	67.28 ± 6.16	70.05 ± 6.10	70.64 ± 6.03
LARCM + Mutual Information LHS	75.92 ± 6.54	77.90 ± 5.54	66.54 ± 6.36	76.90 ± 5.31	79.36 ± 6.39	69.34 ± 6.39	71.08 ± 6.74	71.99 ± 6.34
LARCM + Novelty	81.42 ± 5.70	78.74 ± 5.04	76.15 ± 6.25	77.72 ± 5.01	81.52 ± 5.62	70.02 ± 6.57	71.82 ± 5.81	74.04 ± 6.31
LARCM + Support	84.16 ± 5.29	80.42 ± 5.54	77.45 ± 5.52	79.66 ± 5.21	84.66 ± 5.09	78.08 ± 5.93	79.53 ± 6.18	80.68 ± 5.61

Table A16
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-6.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	88.95 ± 4.94	87.12 ± 4.96	87.99 ± 4.82	83.82 ± 5.33	86.86 ± 5.31	80.83 ± 5.75	78.38 ± 5.76	78.95 ± 6.24
LDA + *bag-of-words*	93.09 ± 4.23	94.62 ± 3.59	92.06 ± 4.69	92.99 ± 3.77	94.12 ± 3.33	90.76 ± 4.44	90.35 ± 4.24	90.78 ± 3.95
LARCM + Added Value	85.49 ± 4.89	88.60 ± 4.20	76.50 ± 5.65	88.23 ± 4.39	88.16 ± 4.62	83.89 ± 4.58	83.29 ± 4.65	83.84 ± 5.00
LARCM + Certainty Factor	87.53 ± 4.57	90.17 ± 4.20	77.07 ± 6.47	89.81 ± 4.13	90.98 ± 3.90	83.41 ± 5.11	84.92 ± 4.91	85.70 ± 4.57
LARCM + Collective Strength	86.43 ± 5.06	88.37 ± 4.59	77.93 ± 5.05	88.47 ± 4.32	89.29 ± 4.30	81.73 ± 5.54	82.96 ± 5.10	83.92 ± 5.28
LARCM + Confidence	86.68 ± 4.56	86.91 ± 4.59	81.57 ± 6.36	85.67 ± 4.77	89.02 ± 4.22	81.25 ± 5.36	82.60 ± 5.14	83.89 ± 4.88
LARCM + Conviction	76.58 ± 6.12	79.12 ± 4.99	72.85 ± 5.60	77.97 ± 5.19	77.73 ± 5.42	73.70 ± 5.38	73.85 ± 5.57	75.18 ± 5.57
LARCM + ϕ-Coefficient	85.88 ± 4.77	90.25 ± 3.95	79.29 ± 5.55	88.83 ± 4.58	90.55 ± 4.04	84.60 ± 5.33	85.68 ± 4.57	85.99 ± 4.86
LARCM + Gini Index	86.96 ± 4.81	88.35 ± 4.28	77.79 ± 6.09	87.51 ± 4.30	88.36 ± 4.08	82.87 ± 5.32	84.44 ± 4.71	83.58 ± 4.74
LARCM + IS	85.76 ± 4.75	87.32 ± 5.04	80.60 ± 4.93	86.59 ± 5.06	87.21 ± 4.89	81.81 ± 5.87	81.38 ± 5.73	81.68 ± 5.80
LARCM + J-Measure	87.58 ± 4.86	86.94 ± 4.75	75.42 ± 6.19	87.94 ± 4.59	89.12 ± 4.36	83.49 ± 4.56	83.48 ± 4.62	84.35 ± 4.66
LARCM + Kappa	87.62 ± 4.33	87.92 ± 4.61	76.91 ± 6.01	88.56 ± 4.57	88.98 ± 4.46	81.95 ± 5.62	82.91 ± 4.81	83.87 ± 5.03
LARCM + Klosgen	85.47 ± 5.10	86.96 ± 4.02	77.76 ± 6.02	86.96 ± 4.02	88.68 ± 4.71	80.58 ± 5.71	81.37 ± 6.02	81.97 ± 5.28
LARCM + Lambda	82.52 ± 4.81	84.57 ± 5.21	75.58 ± 5.92	85.88 ± 5.30	88.31 ± 4.40	78.99 ± 6.17	80.55 ± 5.94	81.55 ± 5.84
LARCM + Laplace	86.00 ± 5.33	88.29 ± 5.06	77.68 ± 5.83	88.95 ± 4.70	90.35 ± 4.18	82.20 ± 4.81	83.71 ± 5.25	84.81 ± 4.83
LARCM + Lift	84.10 ± 5.16	87.02 ± 5.55	78.37 ± 5.44	87.99 ± 4.91	88.95 ± 5.20	79.13 ± 5.83	82.33 ± 5.22	82.59 ± 5.56
LARCM + Mutual Information LHS	87.57 ± 4.14	86.78 ± 4.79	78.60 ± 5.28	88.68 ± 4.73	90.58 ± 4.22	82.46 ± 5.58	84.31 ± 5.31	84.81 ± 5.37
LARCM + Novelty	88.61 ± 4.80	88.76 ± 4.49	78.49 ± 5.85	89.86 ± 4.25	90.93 ± 4.13	84.07 ± 4.77	85.85 ± 5.43	86.34 ± 5.04
LARCM + Support	89.56 ± 4.78	91.00 ± 3.66	84.50 ± 5.01	91.37 ± 3.74	93.22 ± 3.40	87.55 ± 4.07	87.62 ± 4.24	88.39 ± 4.78

Table A17
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-6.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	87.52 ± 4.55	86.13 ± 4.19	79.95 ± 5.43	84.83 ± 4.76	88.50 ± 4.58	74.77 ± 6.87	78.25 ± 6.71	79.34 ± 6.49
LDA + *bag-of-words*	92.27 ± 3.84	92.12 ± 3.71	79.71 ± 5.73	92.08 ± 3.99	92.31 ± 4.11	86.68 ± 4.79	90.00 ± 4.40	90.20 ± 4.56
LARCM + Added Value	86.96 ± 5.52	87.44 ± 4.27	78.42 ± 5.63	89.43 ± 4.37	89.92 ± 4.47	82.07 ± 5.41	83.29 ± 5.26	83.51 ± 4.89
LARCM + Certainty Factor	88.45 ± 4.25	86.81 ± 4.01	83.13 ± 5.79	86.74 ± 4.97	89.53 ± 4.55	82.54 ± 5.76	83.11 ± 5.58	84.32 ± 5.53
LARCM + Collective Strength	87.81 ± 4.93	85.89 ± 4.75	76.87 ± 6.31	86.75 ± 4.99	87.90 ± 4.88	80.12 ± 5.69	80.92 ± 5.55	81.04 ± 4.98
LARCM + Confidence	86.47 ± 4.63	85.74 ± 4.47	80.57 ± 5.98	85.63 ± 5.12	88.28 ± 5.05	79.42 ± 5.64	82.05 ± 5.29	83.13 ± 5.28
LARCM + Conviction	73.45 ± 6.13	74.87 ± 5.97	69.72 ± 6.25	74.26 ± 5.84	75.72 ± 6.51	67.16 ± 6.64	68.83 ± 6.32	70.16 ± 6.15
LARCM + ϕ-Coefficient	87.10 ± 5.36	86.79 ± 4.76	78.65 ± 6.05	85.93 ± 5.20	88.05 ± 4.91	81.76 ± 5.81	82.51 ± 5.18	83.25 ± 5.12
LARCM + Gini Index	87.27 ± 4.64	86.12 ± 4.87	79.59 ± 5.60	85.48 ± 4.75	87.45 ± 4.84	79.54 ± 5.41	80.89 ± 5.86	81.98 ± 5.30
LARCM + IS	87.05 ± 4.61	87.45 ± 4.51	78.17 ± 6.22	85.97 ± 4.94	86.63 ± 5.37	80.46 ± 6.20	81.66 ± 5.44	81.70 ± 5.38
LARCM + J-Measure	88.13 ± 4.65	85.54 ± 4.26	81.32 ± 6.36	85.31 ± 4.55	88.45 ± 4.64	77.30 ± 5.61	78.22 ± 5.14	79.50 ± 4.86
LARCM + Kappa	86.05 ± 5.19	87.58 ± 4.41	79.68 ± 6.05	87.46 ± 4.56	87.42 ± 4.88	79.79 ± 4.84	80.52 ± 4.81	81.25 ± 5.85
LARCM + Klosgen	86.44 ± 4.51	84.85 ± 4.67	79.85 ± 4.85	83.87 ± 4.66	86.96 ± 4.62	77.20 ± 5.95	79.28 ± 6.46	79.05 ± 6.24
LARCM + Lambda	86.89 ± 4.99	84.92 ± 4.09	75.78 ± 6.19	85.89 ± 4.82	88.40 ± 4.93	77.22 ± 5.92	79.11 ± 6.22	79.30 ± 5.45
LARCM + Laplace	87.72 ± 4.43	86.33 ± 4.33	78.29 ± 5.83	86.15 ± 4.38	87.50 ± 4.78	79.61 ± 5.50	80.79 ± 5.77	79.72 ± 5.80
LARCM + Lift	86.10 ± 4.71	85.79 ± 4.34	79.75 ± 5.75	87.66 ± 4.15	88.15 ± 4.07	80.03 ± 5.54	81.17 ± 5.15	81.29 ± 5.39
LARCM + Mutual Information LHS	86.02 ± 4.63	86.43 ± 4.45	80.19 ± 6.13	86.48 ± 4.98	87.16 ± 4.21	78.70 ± 5.76	80.11 ± 5.27	80.76 ± 5.93
LARCM + Novelty	88.18 ± 4.37	88.40 ± 4.30	79.20 ± 6.55	86.81 ± 4.34	89.74 ± 3.87	82.41 ± 6.15	82.54 ± 5.62	83.34 ± 5.37
LARCM + Support	90.66 ± 4.68	88.35 ± 4.39	86.46 ± 5.02	91.10 ± 4.10	92.56 ± 4.09	87.71 ± 4.48	87.79 ± 4.51	87.64 ± 4.25

Table A18
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-6.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	83.33 ± 4.97	82.52 ± 5.40	72.54 ± 6.25	81.26 ± 5.36	84.35 ± 4.81	69.96 ± 7.40	69.20 ± 7.74	66.57 ± 7.83
LDA + *bag-of-words*	86.54 ± 4.18	86.26 ± 4.31	75.10 ± 5.80	81.23 ± 5.22	87.68 ± 4.42	77.68 ± 5.27	75.24 ± 5.50	70.69 ± 6.36
LARCM + Added Value	87.07 ± 4.38	84.17 ± 4.91	80.19 ± 6.26	83.78 ± 5.74	87.99 ± 4.38	78.95 ± 5.71	79.79 ± 5.80	79.59 ± 5.86
LARCM + Certainty Factor	90.03 ± 4.34	86.50 ± 4.17	82.06 ± 5.16	85.67 ± 5.03	88.92 ± 4.28	79.92 ± 5.51	81.12 ± 6.08	81.81 ± 5.79
LARCM + Collective Strength	88.61 ± 4.45	83.93 ± 4.68	79.61 ± 4.90	85.89 ± 4.98	87.86 ± 4.48	80.50 ± 5.94	81.57 ± 6.13	82.25 ± 5.50
LARCM + Confidence	89.14 ± 4.16	84.70 ± 3.83	83.08 ± 6.19	84.21 ± 4.78	88.72 ± 4.25	80.72 ± 4.86	80.74 ± 5.47	80.99 ± 4.90
LARCM + Conviction	71.57 ± 6.54	77.11 ± 6.32	68.53 ± 7.87	73.75 ± 6.10	77.05 ± 6.14	66.30 ± 6.22	68.82 ± 6.35	68.32 ± 6.14
LARCM + ϕ-Coefficient	89.21 ± 4.87	84.88 ± 5.02	78.81 ± 6.12	85.45 ± 4.91	88.72 ± 4.48	78.29 ± 6.21	80.01 ± 5.90	80.78 ± 5.74
LARCM + Gini Index	88.90 ± 4.74	84.69 ± 4.55	80.07 ± 5.90	85.29 ± 5.04	87.11 ± 4.65	79.08 ± 6.14	80.20 ± 5.81	81.07 ± 5.59
LARCM + IS	88.57 ± 4.28	84.89 ± 4.55	74.60 ± 6.25	86.13 ± 5.08	88.11 ± 4.54	82.17 ± 4.68	83.22 ± 5.11	83.02 ± 4.93
LARCM + J-Measure	87.25 ± 4.92	84.33 ± 4.58	81.10 ± 5.19	85.39 ± 5.22	89.98 ± 4.35	79.27 ± 5.17	81.05 ± 5.16	81.58 ± 5.09
LARCM + Kappa	88.94 ± 4.32	83.82 ± 4.69	77.87 ± 5.50	86.06 ± 4.86	87.76 ± 4.84	77.80 ± 6.03	80.32 ± 5.48	80.48 ± 5.06
LARCM + Klosgen	89.15 ± 5.09	85.82 ± 4.68	83.08 ± 5.82	85.01 ± 5.36	88.63 ± 4.79	79.52 ± 6.39	81.28 ± 5.85	81.71 ± 5.50
LARCM + Lambda	86.84 ± 4.69	84.42 ± 4.68	73.96 ± 6.88	83.80 ± 4.60	87.19 ± 4.67	73.59 ± 6.02	76.22 ± 5.43	75.58 ± 6.17
LARCM + Laplace	88.15 ± 4.46	86.38 ± 4.56	80.19 ± 5.18	87.46 ± 4.80	90.16 ± 3.99	80.99 ± 5.63	81.77 ± 5.21	82.39 ± 4.77
LARCM + Lift	87.62 ± 5.18	83.06 ± 4.99	81.06 ± 5.35	83.70 ± 5.24	87.83 ± 4.45	76.11 ± 6.77	76.32 ± 6.14	76.43 ± 5.80
LARCM + Mutual Information LHS	87.91 ± 4.56	85.69 ± 5.34	80.50 ± 5.82	85.64 ± 5.63	88.07 ± 4.68	78.55 ± 6.65	79.94 ± 6.07	80.95 ± 5.74
LARCM + Novelty	89.33 ± 4.29	87.26 ± 4.58	82.22 ± 5.41	87.89 ± 4.51	88.92 ± 4.34	82.96 ± 5.21	85.45 ± 5.15	85.56 ± 4.85
LARCM + Support	90.84 ± 3.52	88.01 ± 4.21	85.93 ± 5.05	89.50 ± 4.30	91.17 ± 3.66	85.31 ± 5.23	86.50 ± 5.22	86.43 ± 5.12

Table A19
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-7.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	86.05 ± 4.85	84.46 ± 5.26	85.93 ± 5.32	81.51 ± 5.58	83.88 ± 5.43	75.93 ± 5.93	75.03 ± 5.54	74.65 ± 5.21
LDA + *bag-of-words*	91.11 ± 4.00	91.40 ± 4.03	87.76 ± 4.52	90.32 ± 4.36	92.13 ± 3.82	87.16 ± 4.59	85.71 ± 4.75	86.61 ± 4.90
LARCM + Added Value	85.22 ± 5.23	88.05 ± 4.50	84.22 ± 5.86	87.37 ± 4.68	88.76 ± 4.33	81.74 ± 5.59	83.02 ± 5.31	82.91 ± 5.26
LARCM + Certainty Factor	87.19 ± 4.52	88.91 ± 3.58	85.21 ± 5.67	87.93 ± 3.73	89.81 ± 3.53	81.22 ± 5.65	81.60 ± 5.41	82.88 ± 4.78
LARCM + Collective Strength	84.75 ± 5.03	86.82 ± 4.19	85.37 ± 4.40	86.71 ± 4.24	87.63 ± 4.27	81.59 ± 5.35	83.52 ± 5.03	84.26 ± 4.64
LARCM + Confidence	85.17 ± 5.36	87.49 ± 4.30	84.95 ± 4.76	86.96 ± 4.42	88.19 ± 4.63	79.41 ± 5.57	82.24 ± 4.94	82.71 ± 5.28
LARCM + Conviction	69.27 ± 6.50	76.80 ± 5.74	73.29 ± 5.94	74.52 ± 5.48	73.95 ± 5.92	70.02 ± 6.15	70.43 ± 6.01	71.15 ± 5.98
LARCM + ϕ-Coefficient	86.74 ± 4.78	88.59 ± 4.40	86.93 ± 5.02	87.99 ± 4.61	88.91 ± 3.99	82.09 ± 5.23	83.96 ± 5.18	84.60 ± 4.63
LARCM + Gini Index	84.76 ± 5.11	87.21 ± 4.35	84.99 ± 4.88	87.02 ± 4.57	88.46 ± 4.17	81.17 ± 4.74	82.41 ± 4.95	82.81 ± 5.03
LARCM + IS	85.48 ± 4.77	87.21 ± 4.45	86.13 ± 4.52	86.42 ± 4.47	88.43 ± 4.70	80.26 ± 5.33	81.58 ± 5.34	81.71 ± 5.66
LARCM + J-Measure	84.55 ± 4.70	87.53 ± 4.63	86.96 ± 5.04	87.10 ± 4.57	88.51 ± 4.36	81.07 ± 5.82	81.30 ± 5.25	82.14 ± 5.40
LARCM + Kappa	82.83 ± 5.00	87.57 ± 4.58	84.11 ± 5.19	86.90 ± 4.92	88.50 ± 4.43	81.28 ± 4.80	83.30 ± 5.11	83.86 ± 5.20
LARCM + Klosgen	87.53 ± 4.50	87.67 ± 4.15	85.48 ± 4.68	86.50 ± 4.49	89.44 ± 4.41	81.94 ± 5.44	81.92 ± 5.29	81.94 ± 5.24
LARCM + Lambda	82.82 ± 5.26	86.12 ± 5.00	81.75 ± 5.98	84.67 ± 5.46	85.87 ± 5.42	77.49 ± 6.18	79.74 ± 5.82	80.23 ± 5.77
LARCM + Laplace	84.69 ± 5.22	87.93 ± 4.98	85.50 ± 5.05	87.50 ± 5.18	88.31 ± 4.55	81.04 ± 5.96	82.84 ± 5.24	83.62 ± 5.60
LARCM + Lift	85.59 ± 5.36	87.31 ± 4.62	84.46 ± 5.51	86.42 ± 4.75	87.74 ± 4.78	80.55 ± 5.68	80.68 ± 5.62	81.85 ± 5.55
LARCM + Mutual Information LHS	84.77 ± 5.94	87.41 ± 5.17	83.98 ± 5.29	87.54 ± 4.90	88.42 ± 5.04	79.16 ± 5.85	80.12 ± 5.78	82.42 ± 5.50
LARCM + Novelty	84.14 ± 5.22	87.25 ± 4.05	85.68 ± 5.08	86.76 ± 4.20	90.04 ± 4.42	81.77 ± 4.93	83.20 ± 4.80	82.99 ± 5.07
LARCM + Support	88.12 ± 4.36	89.53 ± 3.96	88.29 ± 5.17	90.11 ± 4.25	91.47 ± 4.01	87.74 ± 4.99	87.55 ± 4.74	88.23 ± 4.97

Table A20
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-7.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	81.88 ± 5.89	83.24 ± 5.85	83.56 ± 5.69	83.07 ± 5.86	83.22 ± 5.59	72.79 ± 6.74	72.54 ± 6.85	73.75 ± 6.40
LDA + *bag-of-words*	84.73 ± 4.73	87.23 ± 4.42	82.19 ± 6.00	85.44 ± 5.02	87.87 ± 4.61	80.05 ± 5.35	82.01 ± 5.12	80.47 ± 5.44
LARCM + Added Value	84.39 ± 4.58	88.25 ± 4.12	84.28 ± 4.27	86.63 ± 3.95	88.38 ± 3.65	79.00 ± 5.17	79.87 ± 5.66	80.87 ± 5.44
LARCM + Certainty Factor	84.31 ± 5.86	88.68 ± 3.97	86.55 ± 4.04	86.18 ± 4.11	89.04 ± 4.11	79.15 ± 5.86	80.73 ± 6.02	81.99 ± 5.60
LARCM + Collective Strength	85.52 ± 5.02	89.23 ± 4.21	84.22 ± 4.71	87.49 ± 4.61	88.66 ± 4.36	80.21 ± 5.48	81.88 ± 5.12	83.03 ± 5.52
LARCM + Confidence	82.80 ± 5.10	88.31 ± 4.43	86.46 ± 5.32	87.10 ± 4.59	90.02 ± 4.61	79.70 ± 5.60	82.07 ± 5.62	82.56 ± 5.73
LARCM + Conviction	67.86 ± 5.32	76.67 ± 6.04	69.52 ± 6.07	75.16 ± 6.03	75.84 ± 6.05	69.46 ± 6.25	70.59 ± 6.51	72.73 ± 6.40
LARCM + ϕ-Coefficient	86.14 ± 4.62	88.98 ± 3.73	87.49 ± 4.50	88.04 ± 4.13	90.66 ± 3.87	81.52 ± 5.07	83.07 ± 4.83	84.05 ± 4.81
LARCM + Gini Index	86.04 ± 4.84	87.89 ± 4.35	87.31 ± 4.60	87.10 ± 3.89	89.47 ± 3.71	80.41 ± 5.52	81.83 ± 5.16	83.01 ± 5.24
LARCM + IS	85.29 ± 4.75	88.70 ± 4.55	84.20 ± 4.57	86.34 ± 4.98	88.89 ± 4.37	78.38 ± 5.62	79.45 ± 5.46	81.67 ± 5.39
LARCM + J-Measure	83.94 ± 5.23	87.80 ± 4.53	86.99 ± 4.95	86.16 ± 4.58	88.34 ± 3.95	80.00 ± 5.60	81.62 ± 6.02	81.70 ± 5.71
LARCM + Kappa	84.86 ± 5.43	88.19 ± 4.41	88.20 ± 4.30	87.53 ± 4.61	88.96 ± 4.70	80.66 ± 6.09	81.80 ± 5.70	83.69 ± 5.39
LARCM + Klosgen	86.46 ± 4.79	88.98 ± 4.53	86.03 ± 4.48	87.40 ± 4.91	90.49 ± 3.85	79.00 ± 4.95	81.00 ± 4.56	81.92 ± 4.48
LARCM + Lambda	81.94 ± 6.13	87.55 ± 4.79	82.62 ± 5.41	85.25 ± 5.46	86.91 ± 4.90	75.76 ± 6.11	77.69 ± 6.22	79.52 ± 5.86
LARCM + Laplace	84.52 ± 4.62	89.02 ± 4.32	87.59 ± 5.54	87.70 ± 4.08	89.02 ± 4.15	78.21 ± 5.98	80.83 ± 5.56	82.22 ± 5.22
LARCM + Lift	86.31 ± 5.00	88.27 ± 4.72	84.88 ± 4.88	88.06 ± 4.82	89.77 ± 4.15	76.88 ± 6.39	81.24 ± 5.33	82.09 ± 5.30
LARCM + Mutual Information LHS	83.94 ± 5.74	87.38 ± 4.76	82.73 ± 4.56	86.44 ± 4.76	89.49 ± 4.69	77.82 ± 5.48	79.64 ± 5.78	81.00 ± 5.41
LARCM + Novelty	82.86 ± 5.23	87.36 ± 4.42	86.16 ± 5.36	86.74 ± 4.63	88.17 ± 4.34	80.66 ± 5.16	82.05 ± 5.57	82.60 ± 5.55
LARCM + Support	86.57 ± 4.48	90.68 ± 3.71	89.23 ± 4.40	90.49 ± 3.82	91.83 ± 3.89	86.05 ± 4.88	87.37 ± 4.65	87.95 ± 4.84

Table A21
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-7.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	82.37 ± 5.21	78.22 ± 5.52	75.59 ± 6.13	78.81 ± 5.77	81.58 ± 5.37	68.55 ± 5.90	69.34 ± 6.17	66.74 ± 5.76
LDA + *bag-of-words*	83.41 ± 4.69	82.09 ± 4.86	72.50 ± 6.35	81.75 ± 5.06	82.99 ± 5.30	75.78 ± 6.03	72.07 ± 5.64	67.76 ± 6.27
LARCM + Added Value	87.95 ± 4.43	87.71 ± 4.21	85.24 ± 5.38	85.26 ± 4.58	88.27 ± 4.21	79.40 ± 5.72	81.30 ± 5.42	81.53 ± 5.51
LARCM + Certainty Factor	85.39 ± 5.14	87.68 ± 4.29	82.77 ± 5.57	84.03 ± 4.61	88.64 ± 4.02	76.46 ± 4.94	80.90 ± 4.38	82.13 ± 5.10
LARCM + Collective Strength	86.35 ± 5.20	87.33 ± 4.59	85.46 ± 5.10	85.93 ± 5.19	87.87 ± 4.69	79.70 ± 5.09	79.38 ± 6.02	81.19 ± 5.67
LARCM + Confidence	85.02 ± 4.65	87.55 ± 4.36	84.30 ± 4.87	85.33 ± 4.97	89.32 ± 4.12	79.19 ± 5.48	82.13 ± 4.75	83.62 ± 4.85
LARCM + Conviction	66.99 ± 5.94	77.17 ± 5.73	69.78 ± 7.37	75.51 ± 5.38	75.46 ± 5.87	65.74 ± 6.21	68.19 ± 6.53	69.19 ± 6.03
LARCM + ϕ-Coefficient	87.49 ± 4.21	88.17 ± 4.33	87.36 ± 5.22	85.10 ± 4.83	89.30 ± 4.67	80.71 ± 6.32	82.18 ± 5.74	82.94 ± 6.03
LARCM + Gini Index	85.50 ± 5.03	88.46 ± 4.41	84.99 ± 4.30	85.99 ± 4.83	89.15 ± 4.12	81.37 ± 5.11	82.00 ± 5.28	82.73 ± 5.49
LARCM + IS	84.71 ± 5.30	85.46 ± 4.94	82.56 ± 5.61	83.04 ± 5.58	86.79 ± 4.96	78.28 ± 5.33	80.60 ± 4.88	80.52 ± 5.23
LARCM + J-Measure	86.37 ± 4.62	87.12 ± 4.27	84.84 ± 4.95	85.60 ± 4.89	87.82 ± 4.48	79.08 ± 5.87	81.64 ± 5.33	82.39 ± 5.00
LARCM + Kappa	85.31 ± 5.12	87.02 ± 4.59	82.41 ± 4.89	84.89 ± 4.98	87.81 ± 3.90	79.76 ± 5.57	81.56 ± 5.29	81.39 ± 5.20
LARCM + Klosgen	86.50 ± 5.13	88.29 ± 3.52	85.45 ± 5.20	85.01 ± 4.21	88.55 ± 3.68	79.15 ± 5.54	81.03 ± 5.57	82.14 ± 5.22
LARCM + Lambda	84.44 ± 5.40	86.10 ± 4.29	80.79 ± 5.87	83.22 ± 5.30	87.72 ± 4.49	77.55 ± 5.45	78.59 ± 4.44	79.89 ± 5.15
LARCM + Laplace	84.99 ± 4.30	88.55 ± 3.89	84.09 ± 5.26	85.63 ± 4.62	88.46 ± 3.58	79.04 ± 5.18	81.68 ± 4.91	82.33 ± 4.80
LARCM + Lift	86.42 ± 4.78	87.57 ± 5.06	80.88 ± 5.81	85.61 ± 4.78	88.06 ± 4.58	80.13 ± 5.92	82.11 ± 6.27	82.07 ± 6.12
LARCM + Mutual Information LHS	86.14 ± 4.33	87.16 ± 4.45	82.32 ± 5.00	84.58 ± 4.54	88.80 ± 4.45	78.50 ± 5.88	81.64 ± 6.00	83.17 ± 5.22
LARCM + Novelty	84.58 ± 4.57	88.51 ± 4.50	88.08 ± 4.62	85.20 ± 4.54	89.00 ± 4.11	80.34 ± 5.07	82.88 ± 4.54	83.39 ± 4.63
LARCM + Support	88.52 ± 4.57	90.98 ± 3.99	89.62 ± 4.61	89.53 ± 4.04	91.90 ± 3.69	87.08 ± 4.77	88.43 ± 4.55	89.53 ± 4.25

Table A22
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection ACM-8.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	81.63 ± 5.52	81.03 ± 6.03	79.70 ± 5.88	77.42 ± 6.46	81.94 ± 5.89	72.91 ± 6.83	75.37 ± 6.81	73.90 ± 6.47
LDA + *bag-of-words*	86.22 ± 4.80	87.35 ± 4.64	83.32 ± 4.99	86.52 ± 4.57	86.04 ± 4.89	83.28 ± 5.09	83.74 ± 5.10	84.17 ± 5.02
LARCM + Added Value	69.81 ± 6.15	74.49 ± 5.82	62.93 ± 5.84	74.21 ± 6.22	77.12 ± 5.60	66.65 ± 6.27	67.91 ± 6.22	68.37 ± 6.29
LARCM + Certainty Factor	72.26 ± 6.48	74.60 ± 5.91	64.16 ± 7.32	74.26 ± 5.64	76.10 ± 6.02	66.77 ± 5.96	69.60 ± 5.58	70.21 ± 6.06
LARCM + Collective Strength	69.66 ± 6.71	72.36 ± 6.01	60.10 ± 6.25	71.35 ± 6.05	73.31 ± 6.08	64.49 ± 6.87	65.92 ± 7.17	67.35 ± 6.91
LARCM + Confidence	70.86 ± 6.30	74.74 ± 5.53	65.75 ± 6.70	73.63 ± 5.62	76.44 ± 5.98	66.79 ± 6.00	67.74 ± 6.03	67.70 ± 6.33
LARCM + Conviction	55.54 ± 7.05	63.28 ± 7.16	53.80 ± 7.65	61.06 ± 7.10	62.17 ± 6.97	54.72 ± 7.36	56.59 ± 7.24	56.73 ± 6.86
LARCM + ϕ-Coefficient	70.75 ± 6.15	73.01 ± 6.20	64.04 ± 7.07	74.04 ± 6.44	74.34 ± 5.86	65.30 ± 6.57	66.71 ± 7.06	66.49 ± 7.22
LARCM + Gini Index	70.68 ± 5.96	77.75 ± 5.32	65.82 ± 6.72	75.24 ± 5.58	78.07 ± 5.44	67.19 ± 7.70	69.89 ± 6.95	70.66 ± 6.74
LARCM + IS	73.26 ± 5.20	74.04 ± 5.45	63.34 ± 6.57	72.77 ± 5.42	75.10 ± 5.93	66.08 ± 6.31	66.24 ± 5.78	66.29 ± 6.14
LARCM + J-Measure	70.24 ± 6.52	74.83 ± 5.00	62.04 ± 6.45	73.03 ± 6.06	75.05 ± 5.46	65.08 ± 6.31	67.59 ± 6.46	67.84 ± 6.02
LARCM + Kappa	72.08 ± 6.16	75.93 ± 5.48	66.36 ± 6.53	74.30 ± 5.88	77.75 ± 5.72	70.66 ± 5.80	69.31 ± 5.45	68.96 ± 5.63
LARCM + Klosgen	69.22 ± 5.49	74.22 ± 5.99	63.64 ± 6.39	72.84 ± 6.10	76.29 ± 6.11	66.07 ± 6.20	67.11 ± 6.07	66.82 ± 6.14
LARCM + Lambda	68.97 ± 5.63	72.80 ± 6.26	64.41 ± 6.93	70.78 ± 6.23	73.73 ± 6.48	60.22 ± 6.35	61.25 ± 6.79	64.20 ± 6.53
LARCM + Laplace	70.03 ± 5.77	74.34 ± 6.24	59.80 ± 7.09	72.86 ± 5.65	74.34 ± 5.19	62.43 ± 6.36	67.08 ± 6.08	67.45 ± 5.95
LARCM + Lift	71.30 ± 6.24	74.52 ± 5.21	66.89 ± 6.77	73.90 ± 5.09	75.77 ± 5.58	66.76 ± 6.01	70.30 ± 5.39	69.45 ± 5.76
LARCM + Mutual Information LHS	67.67 ± 6.46	73.77 ± 5.23	61.58 ± 6.03	72.77 ± 5.90	75.75 ± 5.84	64.30 ± 6.60	66.57 ± 6.60	66.89 ± 6.65
LARCM + Novelty	72.19 ± 6.50	75.02 ± 6.38	65.72 ± 6.92	74.88 ± 6.74	75.48 ± 5.74	66.76 ± 6.29	67.48 ± 6.70	67.50 ± 6.80
LARCM + Support	76.01 ± 5.86	77.70 ± 5.30	65.41 ± 6.95	79.36 ± 4.97	80.34 ± 4.93	72.55 ± 5.74	74.79 ± 5.31	74.71 ± 5.31

Table A23
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection ACM-8.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	77.82 ± 5.58	77.24 ± 6.04	72.38 ± 6.54	75.50 ± 6.73	80.23 ± 6.38	63.46 ± 6.95	65.94 ± 7.13	69.35 ± 6.49
LDA + *bag-of-words*	84.16 ± 5.04	84.99 ± 5.11	71.92 ± 7.58	82.83 ± 5.49	85.51 ± 4.86	76.20 ± 5.70	76.95 ± 5.74	76.97 ± 5.09
LARCM + Added Value	76.57 ± 5.72	77.31 ± 6.33	67.25 ± 6.71	73.89 ± 5.99	79.33 ± 5.65	65.10 ± 5.88	66.63 ± 6.01	67.07 ± 5.94
LARCM + Certainty Factor	74.86 ± 5.70	75.44 ± 5.51	65.90 ± 6.82	72.62 ± 5.81	78.31 ± 5.36	64.79 ± 5.92	66.72 ± 5.83	68.12 ± 6.00
LARCM + Collective Strength	73.26 ± 5.96	76.56 ± 6.08	65.90 ± 7.27	72.79 ± 6.46	77.58 ± 5.72	65.47 ± 6.86	65.72 ± 6.43	66.86 ± 5.97
LARCM + Confidence	74.07 ± 6.45	76.85 ± 5.48	67.64 ± 6.22	74.87 ± 5.97	80.11 ± 5.11	67.88 ± 6.78	70.89 ± 6.69	71.36 ± 6.20
LARCM + Conviction	51.74 ± 7.16	60.98 ± 6.80	49.63 ± 6.44	58.98 ± 7.04	58.91 ± 6.91	47.82 ± 6.97	50.34 ± 7.46	52.86 ± 7.10
LARCM + ϕ-Coefficient	76.37 ± 5.84	78.58 ± 5.03	69.39 ± 5.62	76.73 ± 5.35	79.40 ± 5.23	65.86 ± 6.21	66.97 ± 6.61	66.99 ± 6.93
LARCM + Gini Index	76.21 ± 6.22	76.94 ± 6.03	65.17 ± 6.52	74.02 ± 6.02	78.36 ± 5.97	64.39 ± 7.15	67.16 ± 7.05	67.46 ± 7.01
LARCM + IS	73.30 ± 5.66	74.61 ± 5.92	65.46 ± 6.73	72.35 ± 6.17	75.71 ± 5.70	66.03 ± 6.61	67.18 ± 6.24	69.12 ± 5.91
LARCM + J-Measure	74.79 ± 6.86	75.29 ± 5.62	67.91 ± 7.07	72.08 ± 5.51	75.34 ± 5.40	62.26 ± 6.51	64.42 ± 6.20	65.14 ± 6.08
LARCM + Kappa	71.70 ± 5.62	76.79 ± 5.50	67.70 ± 6.31	72.98 ± 5.75	77.58 ± 5.83	64.00 ± 7.03	66.69 ± 6.33	67.92 ± 5.53
LARCM + Klosgen	75.05 ± 5.49	78.00 ± 4.90	68.37 ± 6.10	74.16 ± 4.72	77.30 ± 5.42	63.41 ± 6.04	66.27 ± 6.10	67.26 ± 6.01
LARCM + Lambda	71.36 ± 6.05	74.16 ± 5.13	64.02 ± 6.35	71.01 ± 5.97	76.01 ± 5.00	61.16 ± 6.32	63.36 ± 6.24	64.41 ± 6.38
LARCM + Laplace	75.19 ± 5.72	75.70 ± 5.83	67.93 ± 6.31	73.55 ± 5.90	78.04 ± 6.15	65.65 ± 6.28	66.62 ± 7.09	68.01 ± 6.46
LARCM + Lift	73.28 ± 5.68	76.11 ± 5.73	69.08 ± 6.49	73.03 ± 6.29	78.81 ± 4.90	67.00 ± 5.99	66.51 ± 6.10	67.98 ± 6.50
LARCM + Mutual Information LHS	73.16 ± 6.16	75.40 ± 5.42	68.20 ± 5.82	74.04 ± 5.75	78.98 ± 5.88	65.72 ± 6.67	65.91 ± 6.55	67.69 ± 5.81
LARCM + Novelty	73.31 ± 6.36	75.00 ± 6.04	68.29 ± 6.23	71.85 ± 6.38	76.55 ± 6.22	64.09 ± 6.89	67.47 ± 7.92	68.16 ± 7.62
LARCM + Support	78.11 ± 5.13	79.52 ± 4.83	72.36 ± 5.97	77.62 ± 4.90	80.71 ± 5.12	69.28 ± 6.67	70.96 ± 5.57	72.03 ± 5.68

Table A24
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection ACM-8.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	76.16 ± 5.82	76.03 ± 5.69	70.17 ± 7.05	76.05 ± 5.74	76.71 ± 6.07	63.91 ± 5.95	68.09 ± 5.91	68.08 ± 6.69
LDA + *bag-of-words*	74.36 ± 5.69	80.74 ± 4.60	66.37 ± 5.84	75.88 ± 4.58	78.49 ± 5.21	73.74 ± 5.80	75.74 ± 5.71	74.25 ± 5.76
LARCM + Added Value	76.48 ± 5.77	77.37 ± 5.62	69.96 ± 6.22	73.27 ± 6.18	79.07 ± 5.19	66.25 ± 6.20	66.36 ± 6.63	69.01 ± 6.19
LARCM + Certainty Factor	76.00 ± 5.89	77.85 ± 6.02	68.56 ± 6.15	74.92 ± 6.01	79.96 ± 5.54	66.77 ± 6.64	67.76 ± 6.97	67.94 ± 6.70
LARCM + Collective Strength	75.37 ± 6.00	76.67 ± 5.99	66.12 ± 5.41	71.84 ± 6.81	79.37 ± 5.08	65.04 ± 7.16	65.82 ± 6.64	68.28 ± 6.83
LARCM + Confidence	76.22 ± 5.04	78.18 ± 6.09	66.33 ± 6.84	74.39 ± 6.30	81.14 ± 6.06	67.91 ± 6.83	68.87 ± 6.81	69.32 ± 6.70
LARCM + Conviction	52.87 ± 6.31	66.54 ± 6.96	51.43 ± 7.01	64.02 ± 6.74	62.22 ± 7.31	54.96 ± 7.33	56.31 ± 7.06	57.88 ± 7.12
LARCM + ϕ-Coefficient	76.70 ± 5.49	78.25 ± 6.10	71.13 ± 6.14	75.68 ± 6.66	80.68 ± 5.82	68.84 ± 5.58	70.86 ± 5.92	71.63 ± 6.12
LARCM + Gini Index	76.11 ± 5.44	78.62 ± 6.11	68.74 ± 5.64	75.05 ± 6.28	79.57 ± 5.69	66.68 ± 6.38	68.02 ± 6.20	68.10 ± 6.01
LARCM + IS	74.69 ± 6.01	76.42 ± 5.35	69.59 ± 5.46	73.35 ± 5.31	80.14 ± 5.69	68.18 ± 6.31	68.57 ± 6.06	69.68 ± 6.05
LARCM + J-Measure	77.64 ± 6.08	77.23 ± 5.14	67.47 ± 5.97	73.36 ± 5.78	79.81 ± 5.54	61.36 ± 6.79	65.03 ± 6.35	66.58 ± 6.76
LARCM + Kappa	76.18 ± 6.33	76.84 ± 5.64	68.90 ± 5.87	73.71 ± 5.39	79.57 ± 5.31	66.98 ± 6.83	68.31 ± 6.21	69.28 ± 6.27
LARCM + Klosgen	76.03 ± 5.20	75.94 ± 5.18	68.45 ± 5.75	72.24 ± 5.96	80.08 ± 5.07	65.75 ± 5.64	69.01 ± 5.95	68.27 ± 6.04
LARCM + Lambda	72.76 ± 5.38	73.87 ± 5.61	61.32 ± 6.27	69.48 ± 5.80	75.06 ± 5.60	61.73 ± 7.05	64.07 ± 6.71	64.07 ± 6.46
LARCM + Laplace	76.93 ± 5.29	77.49 ± 5.51	69.92 ± 5.74	74.14 ± 5.80	80.32 ± 4.97	66.63 ± 6.66	68.25 ± 6.34	67.52 ± 6.21
LARCM + Lift	77.35 ± 5.68	75.81 ± 5.10	67.08 ± 6.43	72.66 ± 5.79	79.26 ± 4.95	63.63 ± 6.66	65.53 ± 6.89	66.35 ± 6.50
LARCM + Mutual Information LHS	74.97 ± 6.08	76.84 ± 5.62	68.60 ± 7.11	74.01 ± 5.65	78.62 ± 5.33	65.63 ± 5.89	66.02 ± 6.00	67.22 ± 6.05
LARCM + Novelty	77.21 ± 5.76	78.74 ± 5.47	69.84 ± 6.39	75.40 ± 5.63	79.67 ± 5.05	67.35 ± 7.18	66.20 ± 6.44	67.43 ± 5.78
LARCM + Support	82.00 ± 4.70	82.04 ± 5.51	72.47 ± 6.34	78.65 ± 5.59	84.61 ± 5.59	74.48 ± 6.56	75.08 ± 6.54	74.98 ± 5.96

Table A25
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 50$ in the collection Re8.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	75.54 ± 1.46	86.60 ± 1.12	87.59 ± 1.16	88.44 ± 1.23	89.17 ± 1.19	87.07 ± 1.12	87.24 ± 1.19	86.82 ± 1.22
LDA + *bag-of-words*	82.83 ± 1.31	90.10 ± 0.80	90.89 ± 1.01	93.29 ± 0.78	93.56 ± 0.79	92.90 ± 0.75	92.89 ± 0.80	92.73 ± 0.84
LARCM + Added Value	71.28 ± 1.57	82.79 ± 1.16	82.43 ± 1.26	83.88 ± 1.29	84.40 ± 1.22	82.44 ± 1.30	82.81 ± 1.09	82.96 ± 1.12
LARCM + Certainty Factor	69.24 ± 1.58	82.88 ± 1.16	83.00 ± 1.18	83.85 ± 1.36	84.37 ± 1.37	82.51 ± 1.39	82.82 ± 1.33	83.07 ± 1.31
LARCM + Collective Strength	69.82 ± 1.60	82.11 ± 1.09	81.56 ± 1.15	82.66 ± 1.19	83.20 ± 1.21	81.30 ± 1.16	81.84 ± 1.17	81.98 ± 1.19
LARCM + Confidence	68.11 ± 1.71	83.83 ± 1.22	83.39 ± 1.40	84.31 ± 1.31	85.09 ± 1.29	82.48 ± 1.34	83.04 ± 1.37	83.12 ± 1.45
LARCM + Conviction	68.12 ± 1.73	81.53 ± 1.07	81.87 ± 1.20	81.64 ± 1.11	82.59 ± 1.09	81.11 ± 1.17	81.42 ± 1.13	81.36 ± 1.25
LARCM + ϕ-Coefficient	71.49 ± 1.42	83.63 ± 1.18	83.78 ± 1.32	84.97 ± 1.25	85.54 ± 1.20	83.57 ± 1.26	84.10 ± 1.33	84.06 ± 1.35
LARCM + Gini Index	70.05 ± 1.75	83.49 ± 1.19	82.96 ± 1.19	84.09 ± 1.25	84.75 ± 1.21	82.80 ± 1.25	83.29 ± 1.15	83.35 ± 1.23
LARCM + IS	70.76 ± 1.57	82.51 ± 1.26	82.74 ± 1.25	83.22 ± 1.27	83.72 ± 1.21	82.17 ± 1.38	82.72 ± 1.40	82.72 ± 1.43
LARCM + J-Measure	70.89 ± 1.45	83.09 ± 1.18	82.62 ± 1.34	83.80 ± 1.33	84.18 ± 1.29	82.58 ± 1.36	82.88 ± 1.24	83.01 ± 1.22
LARCM + Kappa	70.27 ± 1.56	83.07 ± 1.03	82.98 ± 1.08	83.88 ± 1.19	84.65 ± 1.17	82.38 ± 1.21	82.89 ± 1.26	83.21 ± 1.29
LARCM + Klosgen	69.85 ± 1.53	83.35 ± 1.11	82.96 ± 1.20	84.25 ± 1.18	84.82 ± 1.15	82.85 ± 1.07	83.34 ± 1.18	83.50 ± 1.11
LARCM + Lambda	67.60 ± 1.55	82.24 ± 1.09	82.51 ± 1.35	82.95 ± 1.21	83.40 ± 1.26	82.12 ± 1.34	82.57 ± 1.25	82.66 ± 1.19
LARCM + Laplace	69.66 ± 1.61	83.42 ± 1.19	83.91 ± 1.16	84.46 ± 1.23	85.06 ± 1.17	83.39 ± 1.23	83.56 ± 1.26	83.63 ± 1.23
LARCM + Lift	69.39 ± 1.84	82.85 ± 1.43	82.68 ± 1.36	83.47 ± 1.49	83.96 ± 1.41	82.21 ± 1.24	82.74 ± 1.36	83.01 ± 1.31
LARCM + Mutual Information LHS	70.27 ± 1.54	83.14 ± 1.15	82.77 ± 1.18	83.45 ± 1.13	83.92 ± 1.19	82.51 ± 1.25	82.84 ± 1.35	82.98 ± 1.24
LARCM + Novelty	70.47 ± 1.60	83.29 ± 1.29	83.22 ± 1.26	83.92 ± 1.28	84.32 ± 1.37	82.61 ± 1.44	83.06 ± 1.49	83.29 ± 1.44
LARCM + Support	73.07 ± 2.39	85.89 ± 1.01	85.49 ± 1.12	86.56 ± 1.01	87.61 ± 1.03	85.58 ± 1.21	85.97 ± 1.05	86.12 ± 1.01

Table A26
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 100$ in the collection Re8.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	72.30 ± 1.52	83.56 ± 1.11	84.44 ± 1.20	87.30 ± 1.10	88.55 ± 1.00	84.38 ± 1.16	84.47 ± 1.19	84.24 ± 1.20
LDA + *bag-of-words*	81.34 ± 1.37	90.30 ± 0.85	90.17 ± 1.06	94.44 ± 0.82	94.77 ± 0.71	92.69 ± 0.83	92.95 ± 0.73	92.97 ± 0.75
LARCM + Added Value	71.47 ± 1.45	84.76 ± 0.93	83.47 ± 1.22	85.52 ± 1.06	85.88 ± 1.04	83.51 ± 1.11	83.68 ± 1.20	83.95 ± 1.21
LARCM + Certainty Factor	73.65 ± 1.61	85.13 ± 1.29	83.62 ± 1.52	85.71 ± 1.32	86.30 ± 1.29	83.86 ± 1.31	84.27 ± 1.21	84.14 ± 1.14
LARCM + Collective Strength	72.26 ± 1.59	84.10 ± 1.26	83.21 ± 1.26	84.72 ± 1.32	85.45 ± 1.24	83.09 ± 1.18	83.40 ± 1.12	83.28 ± 1.29
LARCM + Confidence	74.66 ± 1.49	85.45 ± 1.20	83.96 ± 1.26	86.03 ± 1.18	86.86 ± 1.11	84.25 ± 1.09	84.82 ± 1.19	84.81 ± 1.14
LARCM + Conviction	70.66 ± 1.65	83.27 ± 1.05	82.18 ± 1.32	83.11 ± 1.30	84.23 ± 1.29	81.68 ± 1.30	81.92 ± 1.30	81.98 ± 1.32
LARCM + ϕ-Coefficient	76.83 ± 1.34	86.06 ± 1.10	84.27 ± 1.29	86.86 ± 1.04	87.80 ± 0.96	84.58 ± 1.12	85.05 ± 1.05	85.01 ± 1.12
LARCM + Gini Index	72.15 ± 1.56	85.22 ± 1.14	83.59 ± 1.35	85.55 ± 1.13	86.09 ± 1.13	83.51 ± 1.25	83.88 ± 1.25	84.22 ± 1.24
LARCM + IS	72.21 ± 1.32	84.76 ± 1.22	83.17 ± 1.16	85.12 ± 1.28	85.68 ± 1.32	83.61 ± 1.29	83.56 ± 1.22	83.52 ± 1.20
LARCM + J-Measure	71.48 ± 1.63	84.84 ± 1.16	83.19 ± 1.31	85.28 ± 1.28	86.01 ± 1.39	83.52 ± 1.45	84.07 ± 1.36	84.25 ± 1.34
LARCM + Kappa	71.99 ± 1.59	84.76 ± 0.99	83.08 ± 1.24	85.17 ± 1.07	85.65 ± 1.21	83.34 ± 1.22	83.48 ± 1.20	83.68 ± 1.14
LARCM + Klosgen	72.56 ± 1.46	84.54 ± 1.08	83.48 ± 1.18	85.47 ± 1.21	85.97 ± 1.32	83.22 ± 1.33	83.78 ± 1.32	83.78 ± 1.26
LARCM + Lambda	72.64 ± 1.72	83.96 ± 0.97	82.01 ± 1.28	84.66 ± 1.08	84.98 ± 1.10	82.62 ± 1.20	82.88 ± 1.21	83.01 ± 1.23
LARCM + Laplace	74.28 ± 1.71	85.05 ± 1.11	83.94 ± 1.27	85.52 ± 1.26	86.14 ± 1.32	83.85 ± 1.27	84.06 ± 1.25	84.00 ± 1.22
LARCM + Lift	72.90 ± 1.51	85.14 ± 1.15	83.45 ± 1.30	85.43 ± 1.19	85.91 ± 1.26	83.54 ± 1.29	84.11 ± 1.34	84.30 ± 1.24
LARCM + Mutual Information LHS	72.37 ± 1.52	84.42 ± 0.98	82.88 ± 1.39	84.90 ± 1.06	85.48 ± 1.09	83.09 ± 1.22	83.40 ± 1.15	83.60 ± 1.19
LARCM + Novelty	73.92 ± 1.59	85.36 ± 1.12	83.21 ± 1.34	85.62 ± 1.20	86.20 ± 1.18	83.90 ± 1.26	84.18 ± 1.28	84.17 ± 1.29
LARCM + Support	78.50 ± 1.42	87.47 ± 1.06	86.27 ± 1.23	88.72 ± 1.09	89.86 ± 1.05	87.23 ± 1.13	87.40 ± 1.11	87.68 ± 1.12

Table A27
Accuracy and standard deviation for each classifier and representation model, considering the value of $k = 150$ in the collection Re8.

Model	NB	MNB	J48	SMO c= 1	SMO c= 10	IBk 3	IBk 5	IBk 7
LDA + *bag-of-related-words*	68.19 ± 1.70	81.13 ± 1.14	83.74 ± 1.20	87.65 ± 1.15	89.00 ± 1.12	83.12 ± 1.26	83.90 ± 1.18	83.96 ± 1.25
LDA + *bag-of-words*	77.00 ± 1.26	89.10 ± 0.81	90.24 ± 0.97	94.02 ± 0.75	94.92 ± 0.75	91.44 ± 0.77	91.60 ± 0.74	91.78 ± 0.80
LARCM + Added Value	72.22 ± 1.50	85.80 ± 1.17	83.96 ± 1.42	87.03 ± 1.23	87.51 ± 1.18	84.68 ± 1.32	85.11 ± 1.34	85.01 ± 1.40
LARCM + Certainty Factor	72.09 ± 1.59	85.84 ± 1.12	83.62 ± 1.26	86.60 ± 1.13	87.32 ± 1.14	84.36 ± 1.04	84.53 ± 1.12	84.59 ± 1.11
LARCM + Collective Strength	70.47 ± 1.79	84.51 ± 0.98	82.67 ± 1.24	85.41 ± 1.09	86.27 ± 1.08	82.98 ± 1.31	83.44 ± 1.23	83.64 ± 1.17
LARCM + Confidence	74.42 ± 1.42	86.21 ± 1.06	83.97 ± 1.29	87.37 ± 1.12	87.96 ± 0.97	85.02 ± 1.23	85.24 ± 1.27	85.36 ± 1.26
LARCM + Conviction	69.45 ± 1.66	84.81 ± 1.07	82.95 ± 1.30	85.20 ± 1.15	85.51 ± 1.16	83.19 ± 1.24	83.62 ± 1.15	83.55 ± 1.16
LARCM + ϕ-Coefficient	76.43 ± 1.45	86.17 ± 0.95	84.83 ± 1.36	87.39 ± 0.96	88.23 ± 0.94	84.99 ± 1.12	85.38 ± 1.08	85.31 ± 1.08
LARCM + Gini Index	72.77 ± 1.57	86.32 ± 1.09	83.98 ± 1.19	86.79 ± 1.05	87.58 ± 1.08	84.60 ± 1.18	84.96 ± 1.19	84.96 ± 1.24
LARCM + IS	70.42 ± 1.52	85.11 ± 1.28	83.07 ± 1.37	85.81 ± 1.20	86.30 ± 1.17	83.51 ± 1.37	83.92 ± 1.40	83.91 ± 1.34
LARCM + J-Measure	73.14 ± 1.75	85.67 ± 1.08	83.68 ± 1.39	86.41 ± 1.20	86.83 ± 1.15	84.21 ± 1.18	84.52 ± 1.20	84.53 ± 1.19
LARCM + Kappa	71.50 ± 1.62	85.79 ± 1.18	83.47 ± 1.21	86.58 ± 1.14	87.14 ± 1.04	84.12 ± 1.19	84.63 ± 1.17	84.76 ± 1.13
LARCM + Klosgen	72.68 ± 1.55	86.34 ± 1.08	84.10 ± 1.16	86.98 ± 1.11	87.37 ± 1.11	84.65 ± 1.11	85.16 ± 1.06	85.36 ± 1.11
LARCM + Lambda	69.93 ± 1.67	85.07 ± 1.15	83.52 ± 1.42	85.84 ± 1.11	86.47 ± 1.15	83.58 ± 1.18	84.18 ± 1.15	84.55 ± 1.21
LARCM + Laplace	75.45 ± 1.41	86.28 ± 1.15	84.42 ± 1.18	87.32 ± 1.13	88.12 ± 1.10	85.19 ± 1.14	85.54 ± 1.19	85.44 ± 1.18
LARCM + Lift	70.76 ± 1.66	85.36 ± 1.16	83.63 ± 1.22	85.74 ± 1.27	86.41 ± 1.18	83.98 ± 1.34	84.57 ± 1.39	84.49 ± 1.38
LARCM + Mutual Information LHS	71.14 ± 1.58	85.39 ± 1.19	83.57 ± 1.30	85.99 ± 1.16	86.43 ± 1.03	83.70 ± 1.27	84.11 ± 1.29	84.64 ± 1.26
LARCM + Novelty	72.78 ± 1.55	86.07 ± 1.08	83.68 ± 1.20	86.65 ± 1.12	87.05 ± 1.07	84.32 ± 1.22	84.72 ± 1.22	84.73 ± 1.23
LARCM + Support	77.93 ± 1.38	88.03 ± 0.85	85.98 ± 1.00	89.50 ± 0.85	90.20 ± 0.97	87.43 ± 1.06	87.56 ± 1.03	87.76 ± 1.04

References

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems, 23*(1), 103–145.

Aggarwal, C. C., & Zhai, C. (Eds.) (2012). *Mining text data*. Springer.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th international conference on very large data bases*. In *VLDB'94* (pp. 487–499).

Badawi, D., & Altinçay, H. (2014). A novel framework for termset selection and weighting in binary text classification. *Engineering Applications of Artificial Intelligence, 35*, 38–53.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search (2nd edition) (ACM press books)* (2nd). Addison-Wesley Professional.

Batista, G. E. A. P. A., & Silva, D. F. (2009). How k-nearest neighbor parameters afect its performance. In *X argentine symposium on artificial intelligence (asai), Mar del Plata, Argentina* (pp. 95–106).

Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, july 23–26, 2002, edmonton, alberta, canada* (pp. 436–442). ACM.

Billhardt, H., Borrajo, D., & Maojo, V. (2002). A context vector model for information retrieval. *Journal of the American Society for Information Science and Technology, 53*(3), 236–249.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Braga, I. A. (2014). *Stochastic density ratio estimation and its application to feature selection*. Ph.D. thesis. São Carlos, SP, Brasil: Instituto de Ciências Matemáticas e de Computação (ICMC) - USP.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the fourteenth conference on uncertainty in artificial intelligence* (pp. 43–52).

Carvalho, V. O. (2007). *Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado*. Ph.D. thesis. Universidade de São Paulo.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Nips* (pp. 288–296). Curran Associates, Inc.

Chen, C., & Ren, J. (2017). Forum latent dirichlet allocation for user interest discovery. *Knowledge-Based Syst., 126*, 1–7.

Cheng, X., Miao, D., Wang, C., & Cao, L. (2013). Coupled term-term relation analysis for document clustering. In *Ijcnn* (pp. 1–8). IEEE.

Chien, J., Lee, C., & Tan, Z. (2018). Latent dirichlet mixture model. *Neurocomputing, 278*, 12–22.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Deshpande, M., & Karypis, G. (2004). Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems, 22*(1), 143–177.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing, 5*(1), 4–7.

Domingues, M. A., Sundermann, C. V., Barros, F. M. M., Manzato, M. G., Pimentel, M. G. C., & Rezende, S. O. (2015). Applying multi-view based metadata in personalized ranking for recommender systems. In *Proceedings of the 30th annual ACM symposium on applied computing (sac)* (pp. 1105–1107).

Domingues, M. A., Sundermann, C. V., Manzato, M. G., Marcacini, R. M., & Rezende, S. O. (2014). Exploiting text mining techniques for contextual recommendations. In *2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (iat)* (pp. 210–217).

Farahat, A. K., & Kamel, M. S. (2011). Statistical semantics for enhancing document clustering. *Knowledge and Information Systems, 28*(2), 365–393.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira Jr. , W. (2011). Word co-occurrence features for text classification. *Information System, 36*(5), 843–858.

Gao, Y., Xu, Y., Li, Y., & Liu, B. (2013). A two-stage approach for generating topic models. In *Pakdd: 7819* (pp. 221–232). Springer.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management, 59*, 467–483.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations Newsletter, 11*(1), 10–18.

Hariri, N., Mobasher, B., Burke, R., & Zheng, Y. (2011). Context-aware recommendation based on review mining. In *Ijcai '09: Proceedings of the 9th workshop on intelligent techniques for web personalization and recommender systems* (pp. 30–36).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Sigir* (pp. 50–57). New York, NY, USA: ACM.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning, 95*(3), 423–469.

Kalogeratos, A., & Likas, A. (2012). Text document clustering using global term context vectors. *Knowledge and Information Systems, 31*(3), 455–474.

Kefalas, P., & Manolopoulos, Y. (2017). A time-aware spatio-textual recommender system. *Expert Systems with Applications, 78*(C), 396–406.

Keikha, M., Khonsari, A., & Orumchian, F. (2009). Rich document representation and classification: An analysis. *Knowledge-Based Systems, 22*(1), 67–71.

Kim, H. D., Park, D. H., Lu, Y., & Zhai, C. (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology, 49*(1), 1–10.

Lau, J. H., Baldwin, T., & Newman, D. (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing, 10*(3), 10:1–10:14.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the european chapter of the association for computational linguistics*.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th international conference on machine learning (icml'14)* (pp. 1188–1196).

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data (data–Centric systems and applications)* (2nd). Springer.

Manzato, M. G., Domingues, M. A., Marcacini, R. M., & Rezende, S. O. (2014). Improving personalized ranking in recommender systems with topic hierarchies and implicit feedback. In *22nd international conference on pattern recognition (icpr'14)* (pp. 3696–3701).

Marcacini, R. M., Correa, G. N., & Rezende, S. O. (2012). An active learning approach to frequent itemset-based text clustering. In *21st international conference on pattern recognition (icpr'12)* (pp. 3529–3532). IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26: 27th annual conference on neural information processing systems 2013. proceedings of a meeting held december 5–8, 2013, lake tahoe, nevada, united states.* (pp. 3111–3119).

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Momtazi, S. (2018). Unsupervised latent dirichlet allocation for supervised question classification. *Information Processing & Management, 54*(3), 380–393.

Moura, M. F. (2009). *Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico*. São Carlos, SP, Brasil: Doutorado em ciências da computação e matemática computacional.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Hlt* (pp. 100–108). Stroudsburg, PA, USA: ACL.

Nogueira, B. M., Moura, M. F., Conrado, M. S., Rossi, R. G., Marcacini, R. M., & Rezende, S. O. (2008). Winning some of the document preprocessing challenges in a text mining process. In *Sbbd Porto Alegre : SBC* (pp. 10–18). Porto Alegre : SBC.

Panniello, U., & Gorgoglione, M. (2012). Incorporating context into recommender systems: An empirical comparison of context-based approaches. *Eletronic Commerce Research, 12*(1), 1–30.

Porter, M. F. (1997). An algorithm for suffix stripping. *Readings in Information Retrieval*, 313–316.

Pôssas, B., Ziviani, N., Meira Jr. , W., & Ribeiro-Neto, B. (2002). Set-based model: A new approach for information retrieval. In *Sigir* (pp. 230–237). New York, NY, USA: ACM.

Rossi, R. G., Marcacini, R. M., & Rezende, S. O. (2013). Benchmarking text collections for classification and clustering tasks. *Technical Report*. ICMC - USP - São Carlos.

Rossi, R. G., & Rezende, S. O. (2011). Building a topic hierarchy using the bag-of-related-words representation. In *Doceng* (pp. 195–204). New York, NY, USA: ACM.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc.

Santos, F. F., Rezende, S. O., & de Carvalho, V. O. (2014). Identificando o assunto dos documentos em coleções textuais utilizando termos compostos. In *Xi encontro nacional de inteligência artificial e computacional (eniac 2014)* (pp. 550–557).

Searle, S. R. (1971). *Linear models*. New York, NY: J. Wiley.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J., & Spiteri, R. (2007). Document representation and dimension reduction for text clustering. In *Proceedings of the 2007 ieee 23rd international conference on data engineering workshop*. In *ICDEW '07* (pp. 770–779). Washington, DC, USA: IEEE Computer Society.

da Silva Conrado, M., Gutiérrez, V. A. L., & Rezende, S. O. (2012). Evaluation of normalization techniques in text classification for portuguese. In *Proceedings of the 12th international conference on computational science and its applications - ICCSA - part III* (pp. 618–630).

Song, J., Zhang, Y., Duan, K., Hossain, M. S., & Rahman, S. M. M. (2017). TOLA: Topic-oriented learning assistance based on cyber-physical system and big data. *Future Generation Computer Systems, 75*, 200–205.

Tapi, M. N., Bringay, S., Lavergne, C., Mollevi, C., & Opitz, T. (2017). What patients can tell us: Topic analysis for social media on breast cancer. *JMIR Medical Informatics, 5*(3). e23–e23

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*(1), 141–188.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Icml* (pp. 977–984). New York, NY, USA: ACM.

Wong, S. K. M., Ziarko, W., & Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international acm sigir conference on research and development in information retrieval*. In *SIGIR '85* (pp. 18–25). New York, NY, USA: ACM.

Wu, M.-S., Lee, H.-S., & Wang, H.-M. (2010). Exploiting semantic associative information in topic modeling. In *Slt* (pp. 384–388). IEEE.

Xie, M., Yin, H., Wang, H., Xu, F., Chen, W., & Wang, S. (2016). Learning graph-based poi embedding for location-based recommendation. In *Proceedings of the 25th acm international on conference on information and knowledge management*. In *CIKM '16* (pp. 15–24).

Zhang, F., Yuan, N. J., Lian, D., Xie, X., & Ma, W.-Y. (2016). Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. In *KDD '16* (pp. 353–362).

Zhang, W., Yoshida, T., Tang, X., & Wang, Q. (2010). Text clustering using frequent itemsets. *Knowledge-Based Syst., 23*(5), 379–388.

Zhu, D., Fukazawa, Y., Karapetsas, E., & Ota, J. (2012). Intuitive topic discovery by incorporating word-pair's connection into lda. In *Web intelligence* (pp. 303–310). IEEE.

Zoghbi, S., Vulić, I., & Moens, M.-F. (2016). Latent dirichlet allocation for linking user-generated content and e-commerce data. *Information Science, 367*(C), 573–599.