Cognitive Computation
https://doi.org/10.1007/s12559-021-09818-9

Detection of Sociolinguistic Features in Digital Social Networks for the Detection of Communities

Edwin Puertas[1,2] · Luis Gabriel Moreno-Sandoval[2] · Javier Redondo[3] · Jorge Andres Alvarado-Valencia[2] · Alexandra Pomares-Quimbaya[2]

Received: 13 March 2020 / Accepted: 5 January 2021

Abstract
The emergence of digital social networks has transformed society, social groups, and institutions in terms of the communication and expression of their opinions. Determining how language variations allow the detection of communities, together with the relevance of specific vocabulary (proposed by the National Council of Accreditation of Colombia (Consejo Nacional de Acreditación - CNA) to determine the quality evaluation parameters for universities in Colombia) in digital assemblages could lead to a better understanding of their dynamics and social foundations, thus resulting in better communication policies and intervention where necessary. The approach presented in this paper intends to determine what are the semantic spaces (sociolinguistic features) shared by social groups in digital social networks. It includes five layers based on Design Science Research, which are integrated with Natural Language Processing techniques (NLP), Computational Linguistics (CL), and Artificial Intelligence (AI). The approach is validated through a case study wherein the semantic values of a series of "Twitter" institutional accounts belonging to Colombian Universities are analyzed in terms of the 12 quality factors established by CNA. In addition, the topics and the sociolect used by different actors in the university communities are also analyzed. The current approach allows determining the sociolinguistic features of social groups in digital social networks. Its application allows detecting the words or concepts to which each actor of a social group (university) gives more importance in terms of vocabulary.

Keywords Sociolinguistic · Community discovery · Natural language processing · Social networks · Community detection.

Introduction

The public accessibility of digital social networks has made it easier for people to share their personal information, opinions, photos, interests, thoughts, feelings, etc., generating new terabytes of comments and a hundred petabytes of photos and videos every day [1]. This large amount of data creates opportunities and scenarios for analyzing linguistic variation and change, based on a community approach and through the exploration of language concerning society. That is to say, how language is used for communication between different social groups and people in different social situations [2, 3].

Social networks are based on the existence of patterns associated to their socio-cultural, economic and geographic features. These patterns actually permeate time, space, interactions among actors and communities, and digital identity [4, 5]. Because a social network describes the structure of a given community of speakers, it is necessary to determine the distribution of linguistic elements in it. This implies identifying "who speaks," "what language speaks," "to whom they speak," "with whom they speak" and "for what purpose they do it." Just as well, it is necessary to consider the context of who uses a social network, and where and why they use it. A community has been largely viewed as a group of elements with highly interconnected relationships between them. However, human communities are also defined by shared language [6], even if their current ties or

✉ Edwin Puertas
epuertas@javeriana.edu.co; epuerta@utb.edu.co

[1] Faculty of Engineering, Department of Engineering, Universidad Tecnologica de Bolivar, Cartagena, Colombia

[2] Faculty of Engineering, Engineering School, Pontificia Universidad Javeriana, Carrera 7 No. 40-62, Bogota, Colombia

[3] Department of Communication and Language, Pontificia Universidad Javeriana, Cartagena, Colombia

Published online: 26 January 2021

relationships are weak. Furthermore, once a community is detected, a study of their shared language features allows for a better understanding of community dynamics and values (motivations). Such study might allow not only addressing these groups with tailored messages (even for marketing or public health purposes), but also a better understanding of closed communities in need of intervention, such as networks of extreme politics or terrorism [7].

In the same way, a growing interest in analyzing and modeling the social dimension of language has fostered great interest and collaboration among sociolinguistics and computational linguistics researchers. However, there has been no overview of the common and complementary aspects of the two areas [8]. It is known that the language used in social media is described as a dialect, and in fact, the variation of social media language is aligned with social factors such as geography and ethnicity. Likewise, sociolects are specialized vocabularies used by social subgroups defined by common interests or origins [9, 10]. They actually constitute language use similarities characterizing groups of individuals [11]. Furthermore, sociolinguistics describe a sociolect as involving the use of vocabulary with special characteristics such as phonetic (accents) and syntactic characteristics, which are developed within group language according to the frequency of their interactions. Furthermore, sociolinguistics describes a sociolect as involving the use of vocabulary with special phonetic (accents) or syntactic characteristics, which are developed within group language according to the frequency of their interactions. In addition, sociolinguists focus on finding distinct vocabulary features and reconstructing the linguistic image of the world contained in the specific terminology of a group, based on a wide range of lexical and grammatical characteristics [12–14]. The latest trends in semi-supervised learning for social data analysis are expressed by studies embedding finite and infinite communities on graphs [15]. They raise one question: What issues are still to be resolved in identifying specialized vocabulary used in conversations between users of social groups in social networks? The mapping of 21st-century physics social networks and semantic combinations into hyperbolic spaces has demonstrated how dense, centralized collaboration is associated to a reduction in the space of ideas, with the aim of generalizing modern scholarship and science [16]. Furthermore, Balaanand et al. [17] introduces an enhanced graphics-based semi-supervised learning algorithm (EGSLA) to detect fake users by examining user activity over an extended period of time.

In this sense, Cavallari et al.[18] proposes to preserve community structures composed of densely connected nodes. The usefulness of this approach lies on exploring community integration in large dynamic network structures, with the aim of discovering communities based on similarities shared by their components. Meanwhile, Fani et al.[19] use a neural graph embedding based on temporal content similarity, thus capturing social network connections. This allows better performance when there is content or link-based interpolation. Liu et al.[21] propose a dynamic graph-based embedding (DGE) model for recommending relevant users and new items in real-time. It actually captures temporal semantic effects, social relationships and user behavior sequential patterns. This author presents an unsupervised network embedding method for (attributed multiplex network) maximizing both mutual information between local patches of a graph and the global representation of the entire graph [20]. However, there are few available linguistic resources for small communities due to the amount of data that they handle [22].

On the other hand, Cambria [23] show that hybrid approaches to affection computing and sentiment analysis have given very positive results. This approach exploits statistical and knowledge-based techniques to perform tasks such as emotion recognition and polarity detection in texts or multimodal data. This is the case of Sentic LDA [24], which integrates common sense computation in the calculation of word distributions in Latent Dirichlet Allocation (LDA).

For this reason, the main objective of this research was the detection of communities by identifying and segmenting regions of a digital ecosystem (Colombian universities that are concerned with the issue of accreditation). The results show how the sociolect obtained through the study of shared linguistic characteristics allows a better understanding of the dynamics and values of the community. The particular case studied in this research allows access to two important aspects of the problem. Firstly, it facilitates understanding how professors, students or other people related to the digital ecosystem of universities perceive the quality of these institutions. Secondly, the accreditation factor to which a university is particularly related can be cleared through the analysis of Social Networking Site (SNS) comments. The novelty of the article lies in the use of different techniques of Natural Language Processing, Social Network Analysis and Artificial Intelligence for the detection of communities. For this purpose, the current approach breaks down the sociolinguistic features (texts and relationships) identified for each community (university cluster), taking into account the relevance that these words may have in the communities associated with the quality accreditation factors. In this way, the contribution of this method lies on the relevance that the actors of a community give to words in specific contexts. This is done by assessing the word frequency in these social groups, as compared to the one they exhibit in general language. In this way, words become the main community agglomeration criterion.

Methods

The methodological focus of this research is based on the Design Science Research [25]. It involves an expert who designs a sequence of activities that produce an innovative

and useful artifact for a particular problem. The artifact should not only be evaluated to ensure its usefulness for the specified problem, but also contribute to the development of a novel investigative approach. Besides, it must provide a more effective solution to a problem or solve an unsolved one. The proposed approach is applied over three cycles: Rigor, relevance, and design. These cycles should be seen as gears whose movements affect one another. Also, they should guide the creation of the problem-solving artifact and contribute knowledge through theories, models, methods, and techniques.

Taking into account what Gonzalez and Pomares [26] suggest for this type of approach, it is necessary to justify and articulate the relation between rigor and relevance of the research objective through design tests. The processes carried out in this work are closely related to applied research while uniquely supporting the rigor of computer science investigation. Based on the above, we propose the interdisciplinary integration of computational linguistics and artificial intelligence methods and techniques to develop a model for the detection of sociolinguistic characteristics in digital social networks. Considering this, the identification of a social group or community in a digital social network contributes to the relevance cycle. In the design cycle, the identification of the sociolect and distinctive words of the social group in question are carried out through the extraction and processing phases. Finally, in the model's rigor cycle, identification is achieved through topic detection, semantic correlation, and community identification. Based on the above, Fig. 1 shows the model used to obtain sociolects and communities. In the lines that follow, each of these model phases is described.

Identification of Social Groups

This phase involves identifying a social group or community in a digital social network through language affinities. For this purpose, we used the representation proposed by [27], which describes any social platform through seven functional elements. The corresponding tasks of this phase are described in Algorithm 1.

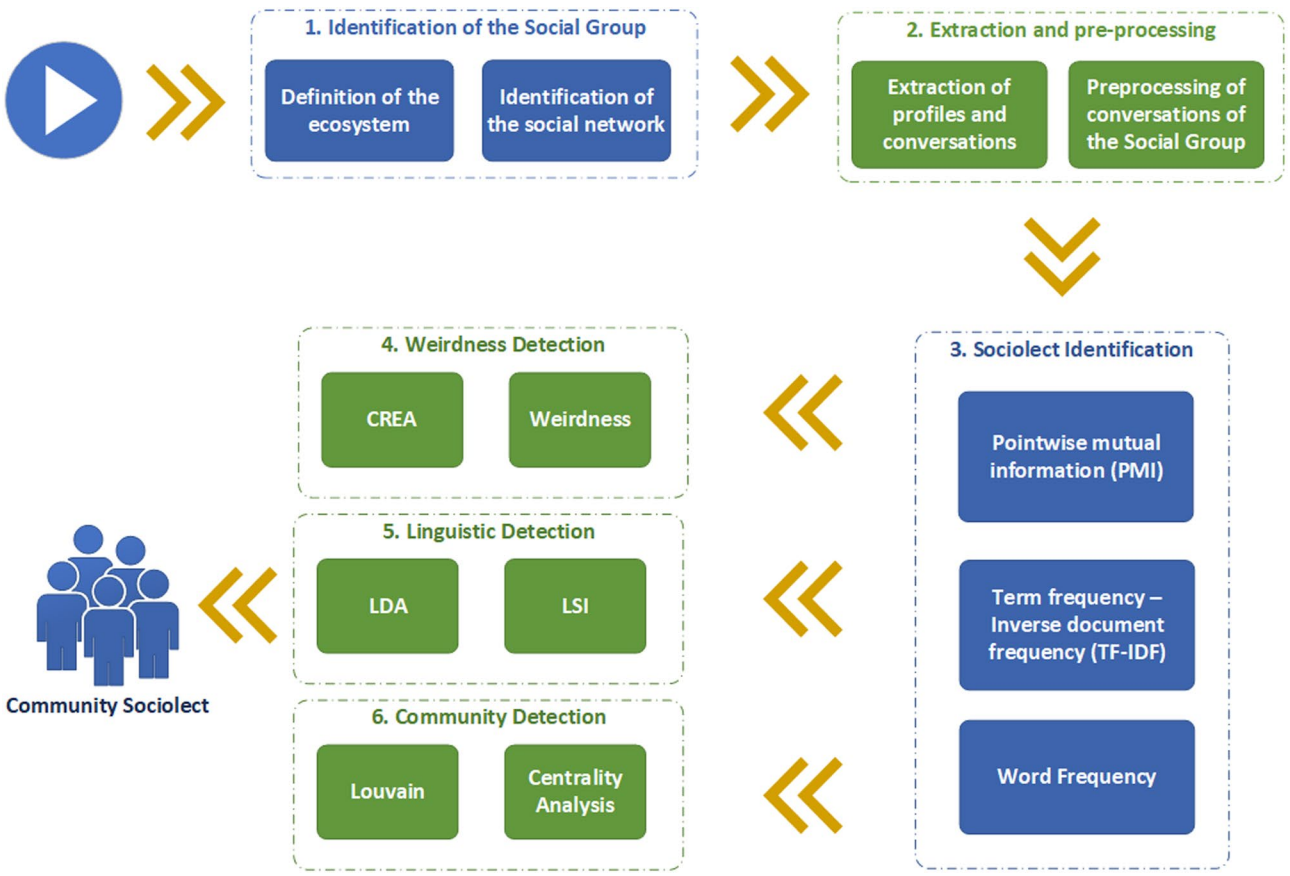

Fig. 1 Model for the detection of sociolinguistic characteristics of in digital social networks

Algorithm 1 Identification of social groups algorithm

A social group is defined as a context-free grammar (G), represented in Equation 1.

$$G = \left\{ v \epsilon \Sigma^* | P \overset{*}{\rightarrow} v \right\} \tag{1}$$

Where:

V is an alphabet of variables (P, I, V, F, S, U).
Σ is an alphabet of terminals (i, v, f, u).
R rule set, which is a finite subset of $V(\Sigma.V\Sigma)$.
P initial symbol, which is an element of V.
$\rightarrow$ indicates that a terminal or unfinished element can be derived from another element of the grammar.
The whole process operates according to the following calculation:

$$P \rightarrow i|iV \tag{2}$$

$$V \rightarrow sSgU \tag{3}$$

$$F \rightarrow s; U \rightarrow u \tag{4}$$

Below, each of the elements of grammar G is described.

P - denotes how a user can determine if other users are accessible on the platform, based on active or inactive status signals.
I - denotes the personal information of the user, which is represented by a finite set of data (e.g., name, age, gender and location) and information (e.g., thoughts, feelings, likes and dislikes).
i $\rightarrow name|age|gender|location|information$
i $\rightarrow i|thoughts|feelings|likes|dislikes$
V - denotes the vocabulary employed by users to communicate on the social network. It is represented by a finite set of alphanumeric characters, together with special characters such as emojis and emoticons.
v $\rightarrow words|phrases|dialect$
F - represents how a user can relate to others in terms of who they follow and who they are followed by.
f $\rightarrow followers|following$
S - is identified by the status of others (including the user) and is represented by a scalar value calculated from the number of favourites, messages, followers and followed posts.
s $\rightarrow active|following$
U - identifies how users interact with one another through language affinity, which is represented by a set of finite interactions between users who share concepts, words, and common interests.
u $\rightarrow users|concepts|words|commoninterests$

Extraction and Processing

Here, information is extracted from the social network associated to the user's profile, considering its followers and conversations. Certain functionalities are used for conversation text analysis, searching for regular, tagger, and tokenized expressions, in order to homogenize the text. Below are the tasks to be performed in order to process the messages extracted from a social network. For such purpose, a message is defined as a set of phrases that express a particular judgment with full meaning and syntactic autonomy. It should be noted that Word Sense Disambiguation (WSD) was not used because the studied Twitter profiles correspond to Colombian Universities and the vocabulary they use corresponds to the same domain. The tasks for this phase are described in Algorithm 2.

Algorithm 2 Preprocessing

Input: Set of messages which can be denoted as:

$$M = \{message_1, message_2 ... message_n\}$$

Sentence: Set of words that express judgment with full meaning and syntactic autonomy.
Tasks:

1. Language detection. This task involves determining the language of the content.
2. Nominal Text. This task extracts a group of words forming a noun phrase, whose core is a noun (noun, pronoun or substantive word). It is noted as in equation 5.
$$N = \{nounPhrase_i \in message\} \tag{5}$$
3. Dependency parsing. This task involves analyzing the dependencies of a sentence, such that they represent its grammatical structure and define relations between words and 'main' words that modify those heads, as shown in equation 6. UD being a finite set of Universal Dependencies.
$$D = \{dependency_i \in UD | dependency_i \in message\} \tag{6}$$
4. Tokenization. This task consists in marking a text string in a sequence of tokens. Where **string** is a finite set of words; **token** is a string of contiguous characters between two spaces or between a space and punctuation, as shown in equation 7.
$$T = \{token : token_i \in message\} \tag{7}$$
5. Stemming. This task consists of reducing the words at their base or root, as shown in equation 8, where Σ is the set of words established in the Current Spanish Reference Corpus *(Corpus de Referencia del Español Actual - CREA)* and is noted as:
$$S = \{word_{root} : word | word_{root} \in \Sigma\} \tag{8}$$
6. Lemmatization. This task gets the flexible form of the words in a sentence and is represented by equation 9, where, A A is the set of words established in the CREA [28] and Σ is the flexible form of a word and is noted as:
$$Lemmatization = \{lemma : word | lemma \in \Sigma, word \epsilon A\} \tag{9}$$
7. Part-of-speech tagging. This task involves automatically extracting some word-related, universal part-of-speech tags [29], as shown in equation 10.
$$W = \{word_i : UniversalPart - of - speechtags | word_i \in message\} \tag{10}$$
8. N-gram Generation. This task involves generating a contiguous sequence of n items from a given sample of text or speech of a given text, as expressed in equation 11:
$$P(w_1^n) = \prod_n^{k=1} P(w_k | w_1^{k-1}) \tag{11}$$
9. NLP. This task involves executing all the previous tasks, where **NLP** is a finite set of Nouns, Dependencies, Tokens, Stemmings, Taglines and Words with Part-of-speech tagging.
10. Clearing text. This task involves the re-tagging of universal texts such as emojis, Uniform Resource Locator (URL), hashtags and mentions. In addition, it removes special characters and stopwords from text.

Output: Dictionary with the result of all the tasks.

Sociolect Identification

In this phase, the frequency of the terms in a collection of documents is calculated using bi-grams and words of more than two characters. Then, all conversations of an account are concatenated by eliminating stopwords, in order to obtain their distinct characteristics. After that, the 100 most relevant features of each account are extracted, and the associated word graphs are drawn. Then, the correlation process must be performed to detect topics associated with the analysis of the sociolect and not only those concerning the most frequently used words by the group of accounts. The tasks for this phase are described in the sociolect identification Algorithm 3.

word frequency of the CREA, which has been prepared by the Spanish Royal Academy [28]was used for such purpose. In the second place, all the frequencies of the words in the general corpus are added up. In the third place, the rate of each word is divided by the sum of the frequencies of all the terms. Finally, the result is multiplied by one million. Eq. 1 describes the formula applied to normalize the words. Eq. 1 describes the formula applied to normalize the words. The equation $WordFreqNorm_{CREA} = \frac{WordFreq_i}{\sum_i^n WordFreqi} \cdot 1,000,000$ describes the formula applied to normalize the words.

Thus, the Weirdness Index (WI) [38] for a particular term is obtained by calculating the quotient between (i) the normalized frequencies of the words found in the corpus

Algorithm 3 Sociolect identification algorithm

Input: Set of messages can be denoted as:

$$M = \{message_1, message_2, ...message_n\}$$

1. Pre-processing Algorithm is invoked.
   (i) Task 10 is executed in order to re-label sociolinguistic features.
   (ii) The equation 10 is applied to extract nominal phrases from the messages and they are aggregated for each message.
   (iii) The equation 11 is applied to extract n-grams from the previous noun phrases.
2. Calculate n-gram frequencies. Can be written as equation 12:

$$TFn - grams_{i,j} = \frac{w_{i,j}^2}{\sum_k w_{i,j}^2} \tag{12}$$

Output: N-gram frequencies.

Weirdness Detection

In this phase, words relating to the categories of an area of interest are identified to determine the most distinctive vocabulary for the users. Likewise, the relationship between the words employed by users and those that are frequently used in the social network is determined. It is necessary to have previously calculated the nominal phrase frequencies of the general and social group corpus. In the first place, rates must be normalized. In the present case, the absolute

of the social group that are greater than one and (ii) the normalized frequency of the words in the corpus CREA. is processed as shown in Eq. 1. The frequency of the term in the corpus CREA is 11.60, and the sum of all frequencies is 134,332,938. For the corpus of the social group, it is 4,000, and the sum is 13.52. Accordingly, the weirdness for the term “estudiante” is described in Eq. 1. The weirdness index represents the importance of the term in context, i.e., the closer the weirdness index comes to zero, the more relevant the value is in context.

$$
\begin{aligned}
FreqCREA_{estudiante} &= 11.60 \\
SumFreq_{CREA} &= 134,332,938 \\
FreqSocialGroup_{estudiante} &= 4,000 \\
SumFreq_{SocialGroup} &= 13.52 \\
FreqNormCREA_{estudiante} &= \frac{11.60}{134,332,938} \cdot 1,000,000 \\
FreqNormCREA_{estudiante} &= 86.36 \\
FreqNormSocialGroup_{estudiante} &= \frac{4,000}{13.52} \cdot 1,000,000 \\
FreqNormSocialGroup_{estudiante} &= 295.79 \\
WI_{estudiante} &= \frac{86.36}{295.79} \\
WI_{estudiante} &= 0.000291962
\end{aligned} \tag{1}
$$

Linguistic Detection

In order to detect social groups, we analyzed the messages posted in the public accounts of Colombian universities and tried to detect communities through their vocabulary. We applied theme-modeling techniques such as LDA, a latent space model used to simplify large text sets and discover hidden themes [24, 31]. This model was used to detect the most relevant words employed in the messages of the social group of Colombian universities. The vocabulary of a community is determined by the ratio of the most ponderous words in their vocabulary. LDA uses statistical methods to asses the relationship between unigrams in texts. This process becomes complex when interpreting the relations between words, because such models do not capture the different meanings of words in context (polysemy). Due to the problems associated with LDA, we resorted to Bringing Bigram to Supervised Topic Model (BL-LDA) [32], which is a supervised generative model for multi-labeled text, that actually extends LDA by applying the bigram concept. Accordingly, this approach can be used as proposed in Algorithm 4.

Algorithm 4 Detection linguistic communities

Input: Set of conversations, where: $M = \{message_1, message_2, ...message_n\}$

1. Pre-processing Algorithm is invoked.
   (i) Task 10 is executed in order to re-label sociolinguistic features.
   (ii) Equation 5 is applied to extract nominal phrases from the messages and they are aggregated for each message
   (iii) Equation 11 is applied to extract bi-grams from the previous noun phrases.
2. Based on the extracted bi-grams, topics are generated by means of the BL-LDA [32]. Equation (1) shows the probability of sampling for topic i with document d, where K is the number of topics, D is the number of documents and N is the number of words in a document. z is the topic associated with the word in the dataset. Ω is the number of tokens. Θ is the distribution of the multinomial topic (Discrete). Φ is the distribution of binomial words (Bernoulli). Λ denotes the presence of tags. η is Dirichlet priori Λ. α is Dirichlet priori Θ. β is Dirichlet priori η. Equation 14 illustrates the mathematical processing of these variables.

$$
F(z_i = j|z_{-i}) \propto \frac{\{N_{w_{i-1},w_i,j}\}_{-i} + \beta_{w_{i-},w_i}}{\{N_j\}_{-i} + \beta} \times \frac{\{N_{d_i,j}\}_{-i} + \alpha_j}{\{N_{d_i}\} + \alpha} \tag{14}
$$

Output: Structural Communities.

Community Detection

In this phase, the Louvain technique is applied to identify communities by analyzing the most central and congruent words among actors of a community with respect to calculated topics, in order to identify the most common ones. Algorithm 5 describes the tasks proposed to identify communities through a structural approach. Nonetheless, Graphics Theory and Network Analysis approaches [33] can be used for this purpose.

Algorithm 5 Detection of structural communities

Input: Set of messages which can be denoted as:

$$M = \{message_1, message_2, ...message_n\} \tag{15}$$

A network is defined as a graph G (V, E), where V is the complete set of nodes and E is the complete set of vertices.

1. For every node i (i ϵ V), Li is denoted by tag i, and N (i) denotes the set of its neighbors.
2. At the beginning of the process, each node is assigned a unique tag (e.g., Li = i).
3. Tags are propagated through the network, each node updating its tag at each iteration to which most of its neighbors share.
4. This process is repeated until each node has one of the most frequent tags of its neighbors, that is, none of the nodes needs to change its tag
5. Finally, communities are built with nodes that share the same tag and can be represented as equation 16.

$$L_i = argmax_l |N^l(i) \tag{16}$$

Output: Structural communities

Results and Discussion

To validate the proposed model for detecting the sociolinguistic characteristics of digital social networks, 113 public and private Colombian universities with institutional Twitter accounts were selected. This choice was based on the hypothesis that social media users share a specific vocabulary defined by their common interests. Then, the sequence of stages proposed in the methodology section are applied.

Identification of Social Groups

The specified social group of universities was selected due to a norm that classifies them with respect to the process of institutional accreditation. According to CNA[34], accreditation is a certification granted by the Ministry of National Education (Ministerio de Educación Nacional - MEN) to universities that achieve higher levels of quality than those required for the operation of the same institutions or their academic programs. It should be noted that institutional accreditation is a voluntary process for institutions and academic programs that, according to the results of an external evaluation, prove the high quality of their service.

To identify vocabulary related to each one of the directives defined to achieve accreditation, we based our hypothesis on the existence of conglomerates of universities using a common vocabulary that concerns a specific quality factor. The directives are grouped into approximately 12 quality factors: Mission and institutional projects, students, professors, academic processes, national and international visibility, research and artistic creation, relevance and social impact, self-assessment and self-regulation, institutional well being, organization, administration and management, academic support resources, and financial resources. A group of experts on the university domain, with doctoral studies in education, and specialized on the quality of higher education were asked to select keywords about the institutional accreditation directives. Table 1 describes the vocabulary used by Colombian Universities with regard to each of the accreditation factors.

Table 1 CNA specialized vocabulary by Quality Factors

Quality factor	Vocabulary
1. Mission and institutional projects	Mission, principles, objectives, nature, image, community.
2. Students	Duties, rights, decision, organs, regulation. permanence, representatives, admission, dropout, exchange, credits, scholarships, subsidies, incentives.
3. Professors	Disciplinary ladder, link, level, degrees, experience, hiring, teaching, assignment, salary, teachers.
4. Academic processes	Flexibility, integrality, interdisciplinarity, science, technology, innovation, culture, curriculum, study plan, TICs , evaluation, relevance, pertinence, profiles.
5. National and international visibility	Internationalization, conventions, cooperation, qualifications, alliances, mobility, visitors, approval, networks, internship.
6. Research, artistic and cultural creation	Scientific knowledge, research, systematization, products, publications, indexing, patents, creation, doctoral dissertation.
7. Pertinence and social impact	Environment, projection, extension, external, interaction, context, regional, local, transfer, graduates, alumni, labor.
8. Self-assessment and self-regulation process	Self-assessment, improvement, quality, control, monitoring, indicators, information, systems.
9. Organization, Administration and Management	Management, organization, administration, performance, documentation, communication, connectivity, leadership, files, managers.
10. Academic support	Resources and physical infrastructure Resources, libraries, laboratories, equipment, audiovisual, didactic, practice, virtual, infrastructure, information technology, offices, bathrooms, cafeterias, green areas, health, security, spaces, conservation.
11. Institutional	Well-being environment, psycho-social, medical services, vulnerable social groups, infrastructure for the disabled, health, emergencies, conflicts, sports.
12. Financial resources	Financial, investment, budget, debt, execution, stability, audit, sustainability, taxes, transparency, reinvestment.

Extraction and Processing

For the extraction sub-phase, the public Twitter application programming interface (API) was used based on three of its functions. The first function is associated with user profiles. The second one has to do with followers, and the third one registers tweets from specific accounts. Twitter information is extracted from a group of previously selected accounts by means of "Phyton." This tool allows extracting information through an API developed in the programming language of a group previously selected Twitter accounts. These accounts are defined as the digital ecosystem or social group. In this study, a mean sociometric homophily assumption about the group was applied to define it, based on its interrelated social variables. In this particular case, the group having digital social network accounts was the object of analysis. Twitter's API was used to extract user profiles, tweets, followers, and mentions from 113 Twitter institutional accounts of Colombian universities. Table 2 details the information collected from the social accounts. Table 3 divides the universities in accredited and non-accredited (public, private) institutions.

Table 2 Collection of data

Features	#
Profiles	54,463
Tweets	517,514
Hashtag	265,715
Follower	54,353
Mention	3,217,660

Sociolect Identification

For this phase, the technique Term Frequency-Inverse Document Frequency (TF-IDF) was used to assess the values assigned to each word included in a message. These were calculated as an inverse proportion of the frequency of each word in a particular message. Words with high TF-IDF values are attributed a strong relationship with the document in which they appear. This, in turn, suggests that if that word appeared in a query, the document could be of interest to the user [14]. This can be expressed formally in the following manner: a set of documents D with the user entering a query q = $\{w_1, w_2, w_3 ... w_n\}$ for a sequence of words w_i.Then, we return a subset $D*$ of D such that for each $d \in D*$, the following probability is maximized: $P(d \mid q, D)$. TF-IDF is a simple and efficient system. The algorithm allows joining words in a query through relevant documents [35]. Figure 2 presents our TF-IDF analysis of the most common keywords

Table 3 Dataset distribution

Category	Public	Private
Accredited	17	23
Not Accredited	16	57

in the corpus in question, with the objective of identifying key themes, trends, and opinions. Figure 3 shows the most significant concepts and the universities that discuss them.

Additionally, the process of vectorization and correlation is performed in this phase. The vectorization of the documents in question is performed through the "CountVectorizer" and "Scikit-learn" libraries, which facilitate vocabulary identification and bigram frequency assessment in each document. This is followed by the implementation of an association measure commonly applied in information statistics, which refers to individual events via the Pointwise Mutual Information (PMI) technique. This tool actually detects topics associated with the sociolect analysis and not only those words most frequently used by this group of people [36]. PMI considers a symmetric matrix with each row/column indexed by a word. The input for (w,w') is $PMI(w, w') = log\frac{p(w,w')}{p(w)\cdot p(w')}$, where p(w,w') is the empirical probability of words w, w' appearing in a window of a certain size in the corpus, and p(w) is the marginal probability of w [37].

For the semantic correlation process, a manually tagged lexicon of words associated to Colombian academic accreditation factors was used. Provided that our hypothesis states that the accounts under analysis are of the official type, we wished to determine which of the studied messages were associated to the 12 factors described in Table 1 as related to disclosure of accreditation processes. Figure 4 shows the 10 most relevant accreditation factors as they relate to the five main cities of

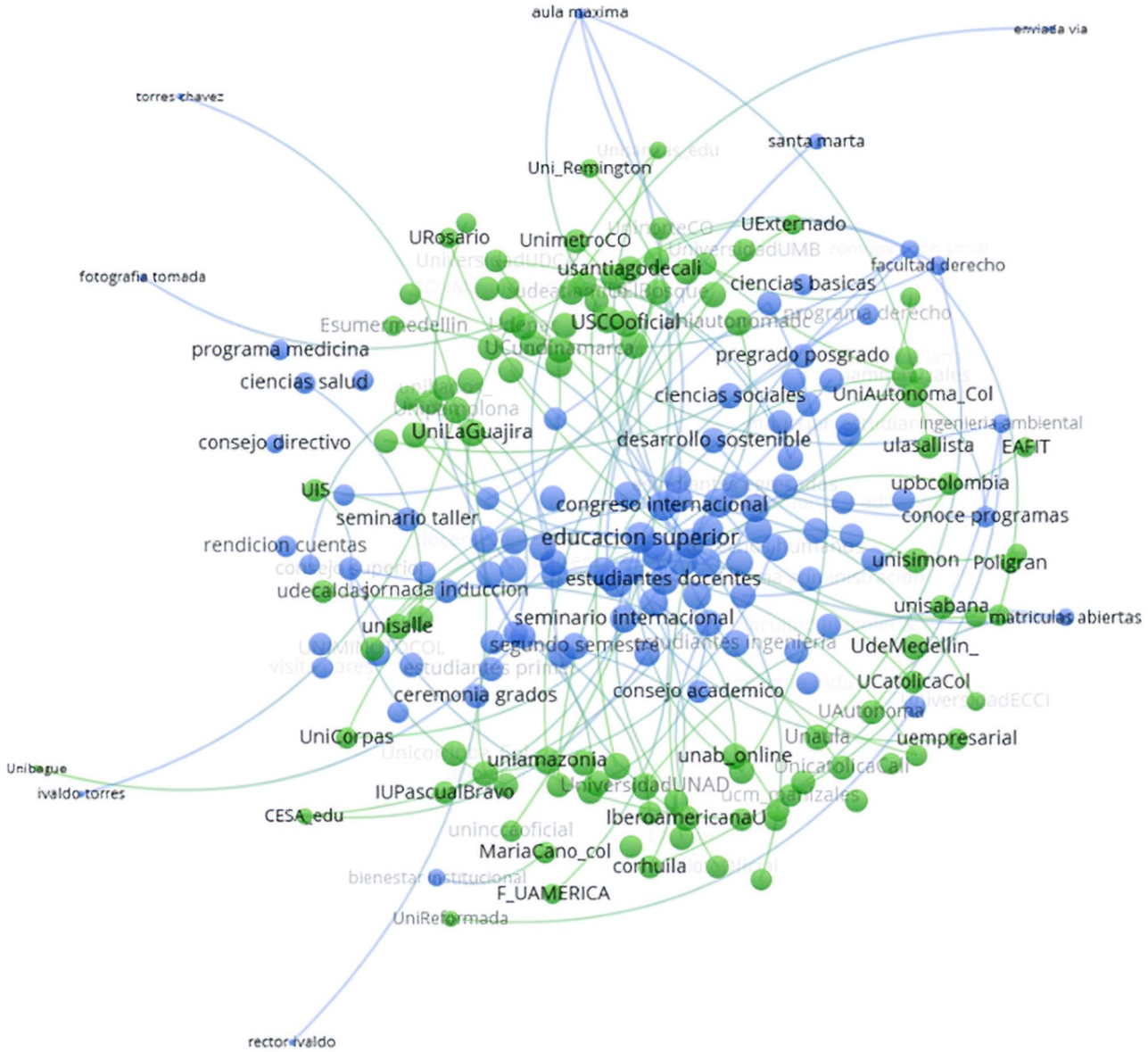

Fig. 2 TF-IDF words/universities

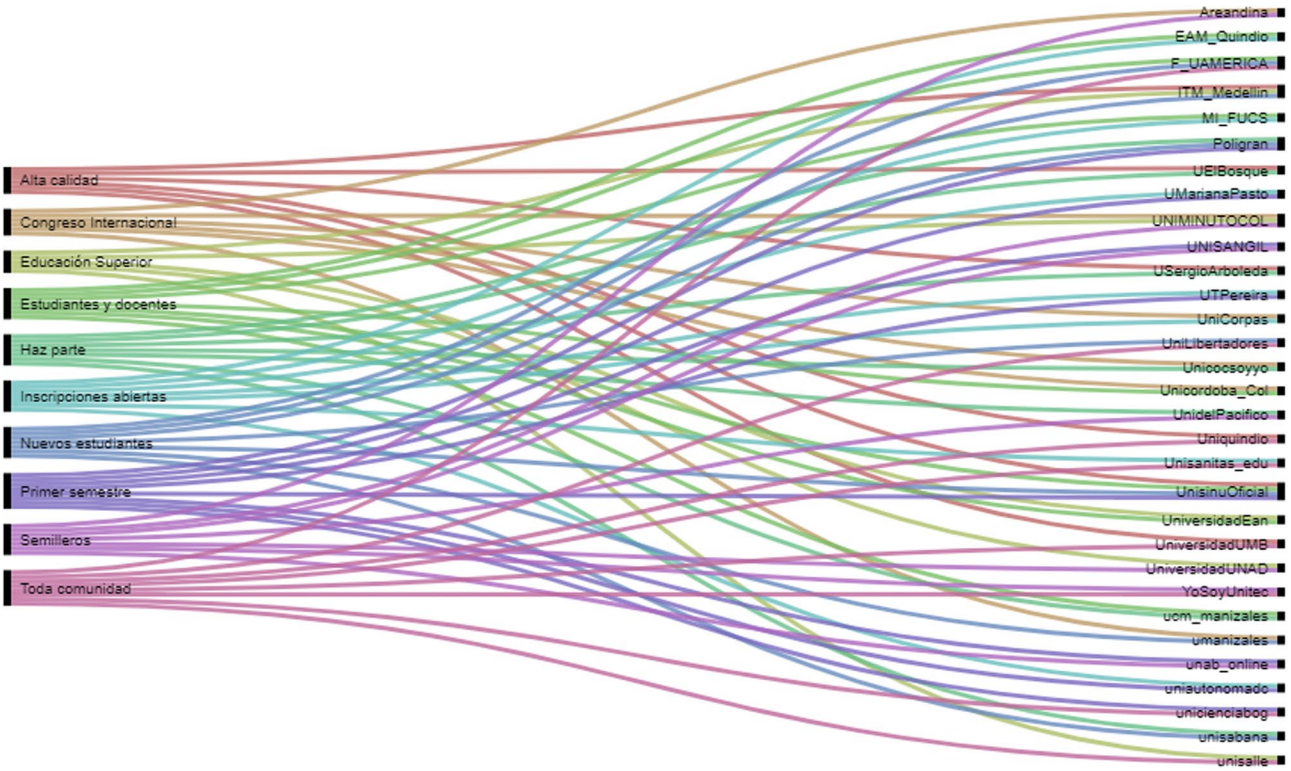

Fig. 3 Concepts/universities

Colombia. “Financial Resources”, “Academic Processes” or “Mission and Institutional Projects” quality factors are not include hereafter, since no university was observed to be related to them.

Weirdness Detection

Weirdness detection was based on the general prevalence of words considered distinct with respect to the general use of words in Spanish as proposed in the *CREA* [28]. The technique Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER) was used to calculate rarity values [30] shows the most distinctive words in larger fonts, and more common words in smaller fonts. In Fig. 5, circles are used to represent inter-topic relations which, depending on their strength (weakness) are depicted closer to (further from) the center of each circle. Additionally, in the bar diagram provided in Fig. 6, 7, the ten most relevant topics are

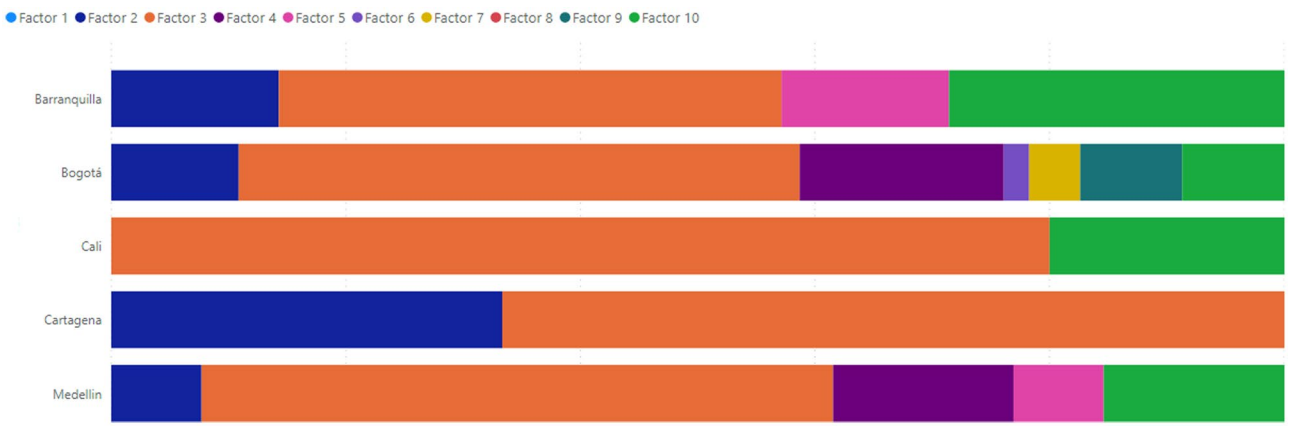

Fig. 4 Top 5 universities by accreditation factors

analyzed. The terms shown in the diagram are the most useful ones to define each topic. The relevance (colored red) of the selected topic is observed in the context of the analyzed corpus (colored blue). The size of the red bar represents the frequency of the term in question, while the blue bar represents the frequency of all other terms in the corpus.

Linguistics Detection

In this phase, themes are associated in a semantic context wherein the most relevant groups of words in the selected sociolect are determined, thus allowing the observation of the accreditation factors. The concepts or quality factors are inferred from the messages of each university through BL-LDA, based on the most frequent words related to the topics extracted from the university corpus. In this way, the accreditation factors are understood as university-specific progress fields. For example, La Salle University can be said to be related to the concept "Higher education," because of the way they use the vocabulary related to that quality factor as it is described in Table 1. In turn, National University of Colombia and Francisco José de Caldas District University are associated to the Superior Council. College of Higher Administration Studies and ICESI University are related to High Degrees of Quality. Del Rosario University and University Sergio Arboleda are related to Seedbed Research. Andes University and Pontifical Xavierian University are related to University Wellbeing. Figure 8 shows the relevance of quality factors by university. In addition, Fig. 8 shows that some universities are in the same group because they focus their efforts on the same quality factors or progress area. This is also related to the type of university. For example, private universities put a lot of effort on issues related to new students, while public universities focus on issues related to student well-being.

Community Detection

The Louvain method was applied to the studied Colombian university Tweets, in order to identify the topics covered by each group. This resulted in 10 clusters that can be observed in Fig. 9. The strength of inter-group relations can be inferred from the colors of the arcs, which allow visualizing specific inter-node interactions across groups. On the other hand, three isolated clusters corresponding to Atlantico University, Pamplona University, and Ibague University can be observed. In turn, Table 4 illustrates the relative importance of each cluster resulting from the analysis of communities in terms of sociolects. This percentage, which we have termed "modularity," measures the relative density of the borders within the communities with respect to the borders outside the communities. The table shows that the most outstanding clusters are 1, 4, and 8, with modularity percentages of 22.4%, 15.12%, and 15.12%, respectively. Clusters 10, 9, and 2 are less critical, scoring 0.98%, 1.46%, and 1.95%, respectively. Representative nodes from each cluster denote official university accounts, except for cluster 1.

In addition, the centrality analysis of the 10 first public university accounts is summarized in Table 5 and Fig. 10. The centrality indicators of the Colombian university

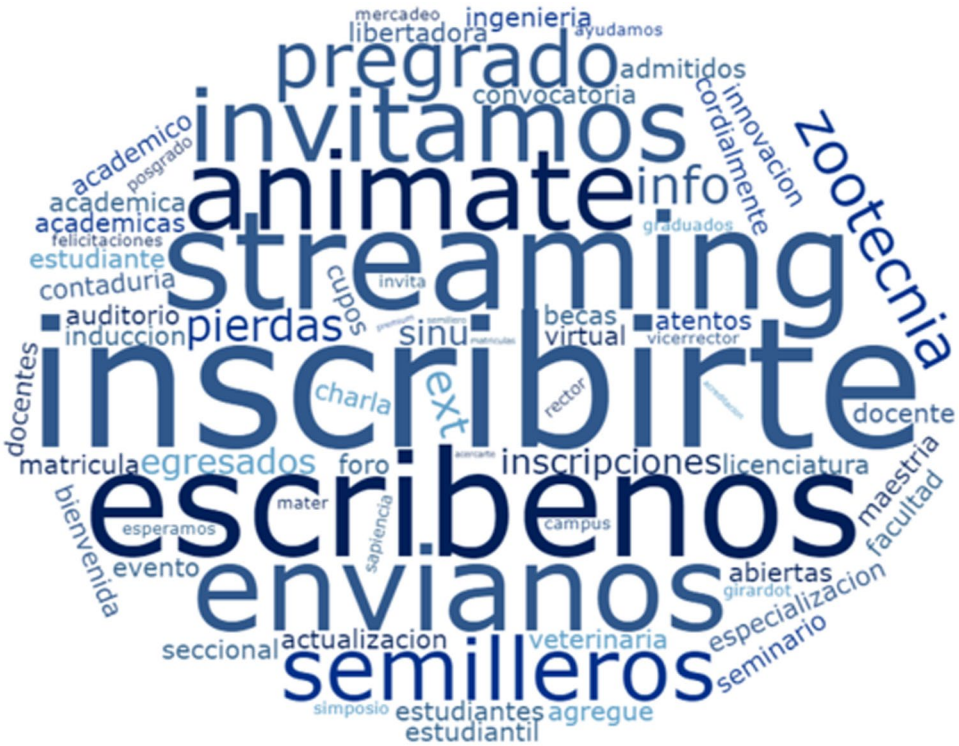

Fig. 5 Distinctive words used in the university community

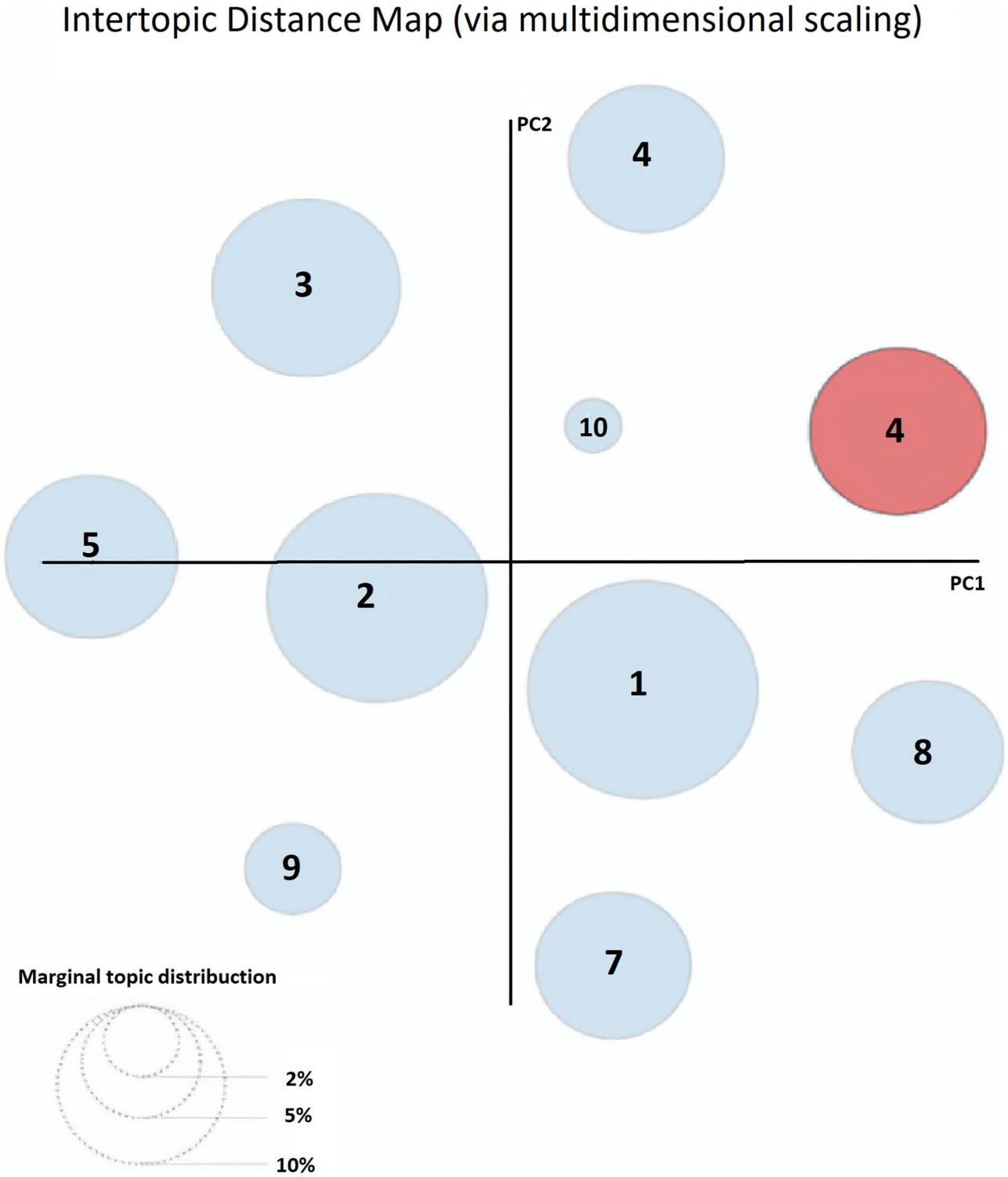

Fig. 6 Relevant words in the university community

community identify public universities such as the University of La Guajira (@Uniguajira), South Colombian University (@US-COo cial), and the University of Pamplona (@ Unipamplona).

In the ecosystem of accredited universities, Fig. 10 emphasizes the importance of public universities, since they receive resources from the Ministry of Education for the accreditation process, while private institutions

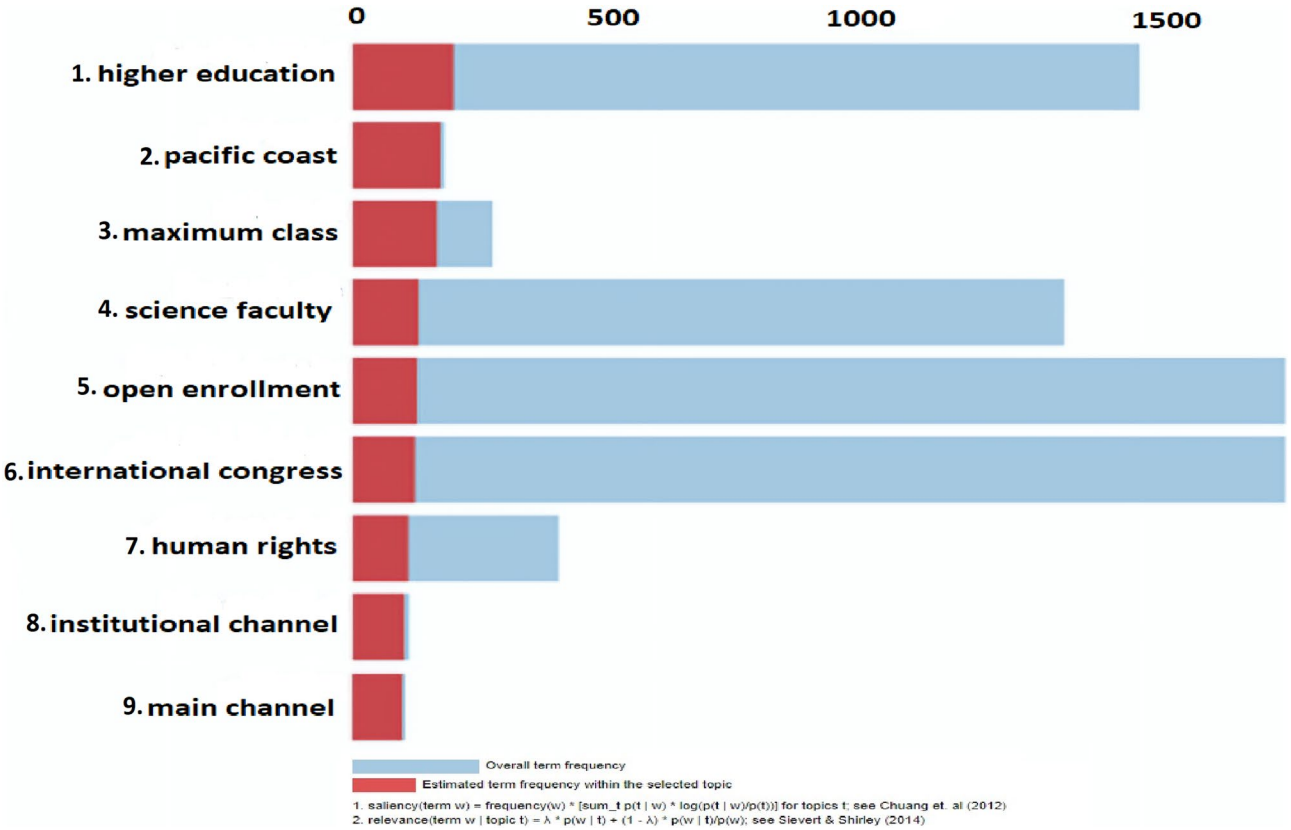

Fig. 7 Relevant words in the university community

must seek their own resources for it. Finally, the structural approach analyses the centrality of words within the corpus of accredited universities. The aim is to find the most relevant bi-grams and establish the sociolect of the group of accredited universities. Figure 11 shows that the most relevant words are those associated to larger nodes. Just as

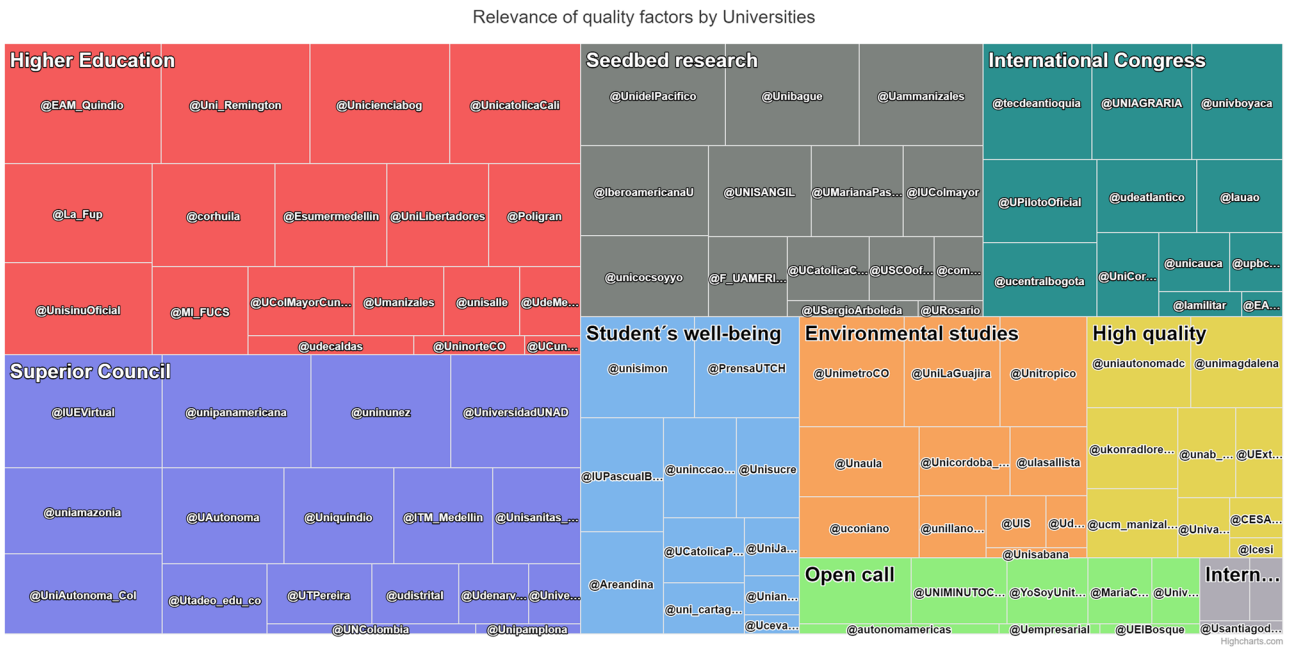

Fig. 8 Relevance by university concepts

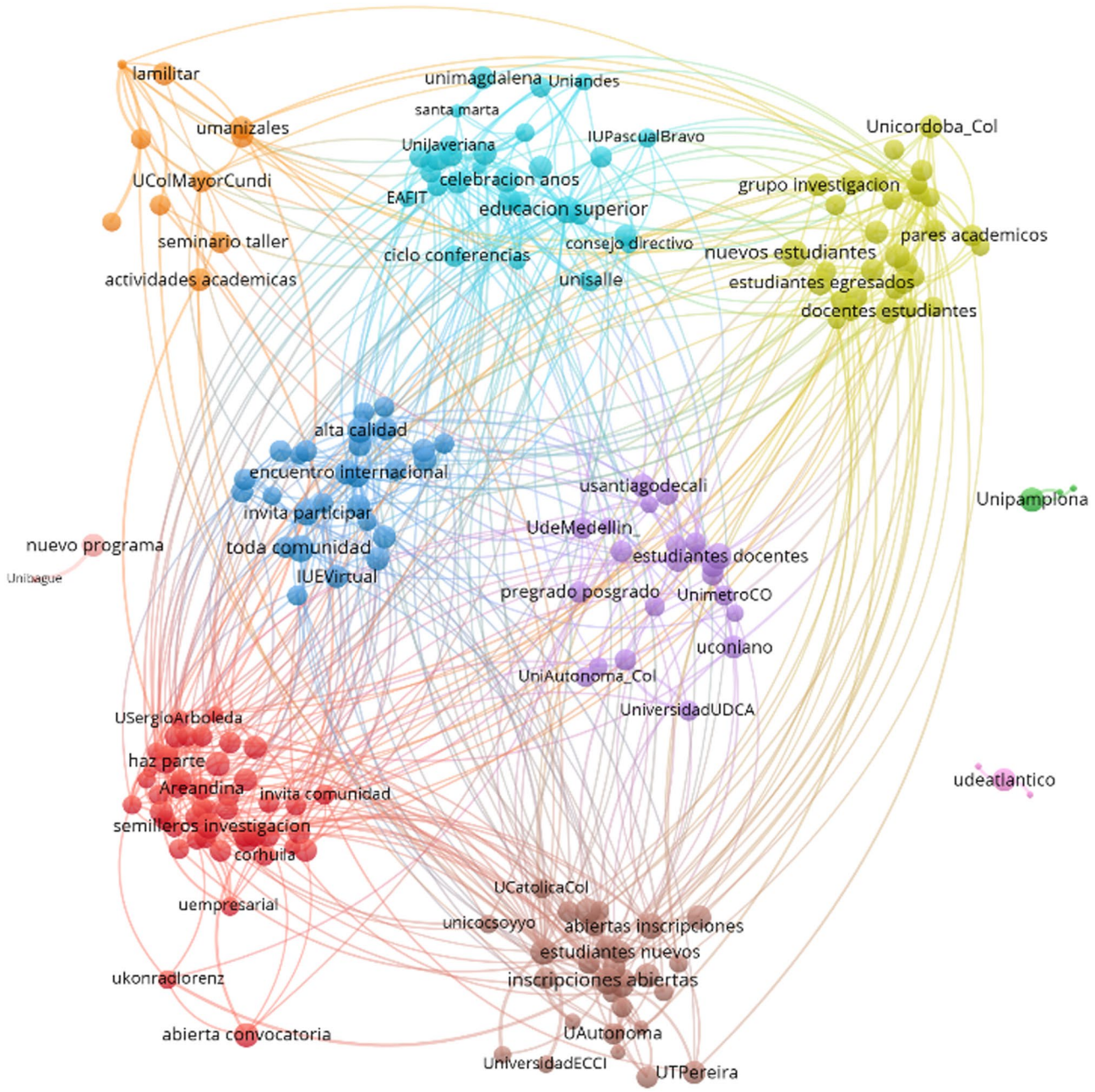

Fig. 9 Graph of university communities

well, it can be observed that some of the bi-grams shown in Fig. 11 are related to the quality factors and vocabulary introduced in Table 1.

The aforementioned results indicate that the current model is useful to determine how Colombian Universities are affected by the diffusion of their quality factors in digital social networks. This, in turn, allows identifying the correlation between the semantic values of the studied public universities, using the quality accreditation vocabulary determined by the CNA. In addition, the results make it possible to determine the aspects to which each university gives more importance concerning the minimum and maximum quality assessment parameters proposed by the CNA. On the other hand, it can be observed how linguistic innovation is spread in society and how new social environments help to determine word correlations by establishing a reference vocabulary. Moreover, the structural approach of the model facilitated the interpretation of the language of the

Table 4 Analysis of communities

Cluster	Freq	Freq%	CumFreq	CumFreq%	Representative
1	46	22.44	46	22.40	All universities
2	4	1.95	50	24.40	University of Pamplona.
3	27	13.17	77	37.60	Central University of Valle del Cauca.
4	31	15.12	108	52.70	University of Cundinamarca.
5	22	10.73	130	63.40	University of Santiago de Cali.
6	29	14.15	159	77.60	University of Los Andes.
7	10	4.88	169	82.40	Nueva Granada Military University.
8	31	15.12	200	97.60	Technological University of Pereira.
9	3	1.46	203	99	University of Atlantico.
10	2	0.98	205	100	University of Ibague.

studied institutions, taking into account their specific social factors and the vocabulary associated to the quality factors in question.

Discussion

In the context of universities, quality processes are monitored by the CNA, which provides basic guidelines for developing high-quality processes and establishes objective metrics to evaluate universities at the regional, national and international scales. Hence, by understanding the multiple benefits of social networks, universities can use them to comment on any topic revolving around quality factors in response to different interlocutors. Since university Tweets can be processed, classified and grouped, the present research study identified comments concerning accreditation factors issued by different interlocutors in a digital academic ecosystem. Since these comments were framed in the institutional dialogue on accreditation factors, they were used to determine the semantic values they contained as proposed by universities, to be contrasted with those employed by the CNA. Concerning the above, the current approach determines the semantic values that a community gives to social network texts concerning the quality factors defined by the CNA. The most relevant bi-grams identified in connection to the universities under study allowed associating the accreditation factors to specific university communities featured by shared sociolinguistic characteristics. This made it possible to determine which evaluation parameters are given more importance by the studied institutions. We found that "Higher Education" is related to colleges such as the National University of Colombia and EAN University. "High Quality" is related to Sergio Arboleda University, Central University of Bogota, and the University of Quindío. "Research Seedbeds" is associated to the Catholic University of Cali, Andean Area and University of the Pacific. Finally, "University Wellbeing" is related to Javeriana University, University of Los Andes, and INCCA University, among others.

Table 5 Analysis of communities

Rank	Vertex	Value	University
1	94	81	University of La Guajira.
2	36	81	South Colombian University.
3	2	80	University of Pamplona.
4	79	78	National University of Distance Education.
5	74	78	University of the Amazon.
6	25	77	University of Nariño.
7	5	77	University of Cundinamarca.
8	22	77	Nueva Granada Military University.
9	80	76	Cooperative University of San Gil.
10	77	76	University of Magdalena.

With respect to the twelve factors of accreditation, it was found that student-focused universities include the San Martín Foundation, University Los Libertadores Foundation, Monserrate University and Unitec. The institutions associated to "Professors" include District University Francisco Jose de Caldas, ICESI, the University of Antioquia, and the Industrial University of Santander. Among those related to "National and International Visibility," we can count UDCA, Unipanamericana, Javeriana University, and the National University of Colombia, while no institution is related to "Financial Resources," "Academic Processes," or "Mission and Institutional Projects."

The most relevant universities in terms of centrality are public ones: The University of La Guajira, South Colombian University, and the University of Pamplona. This highlights the importance of public universities to the analyzed setting, since the Cooperative University of San Gil is the only private one ranked among the top 10 universities.

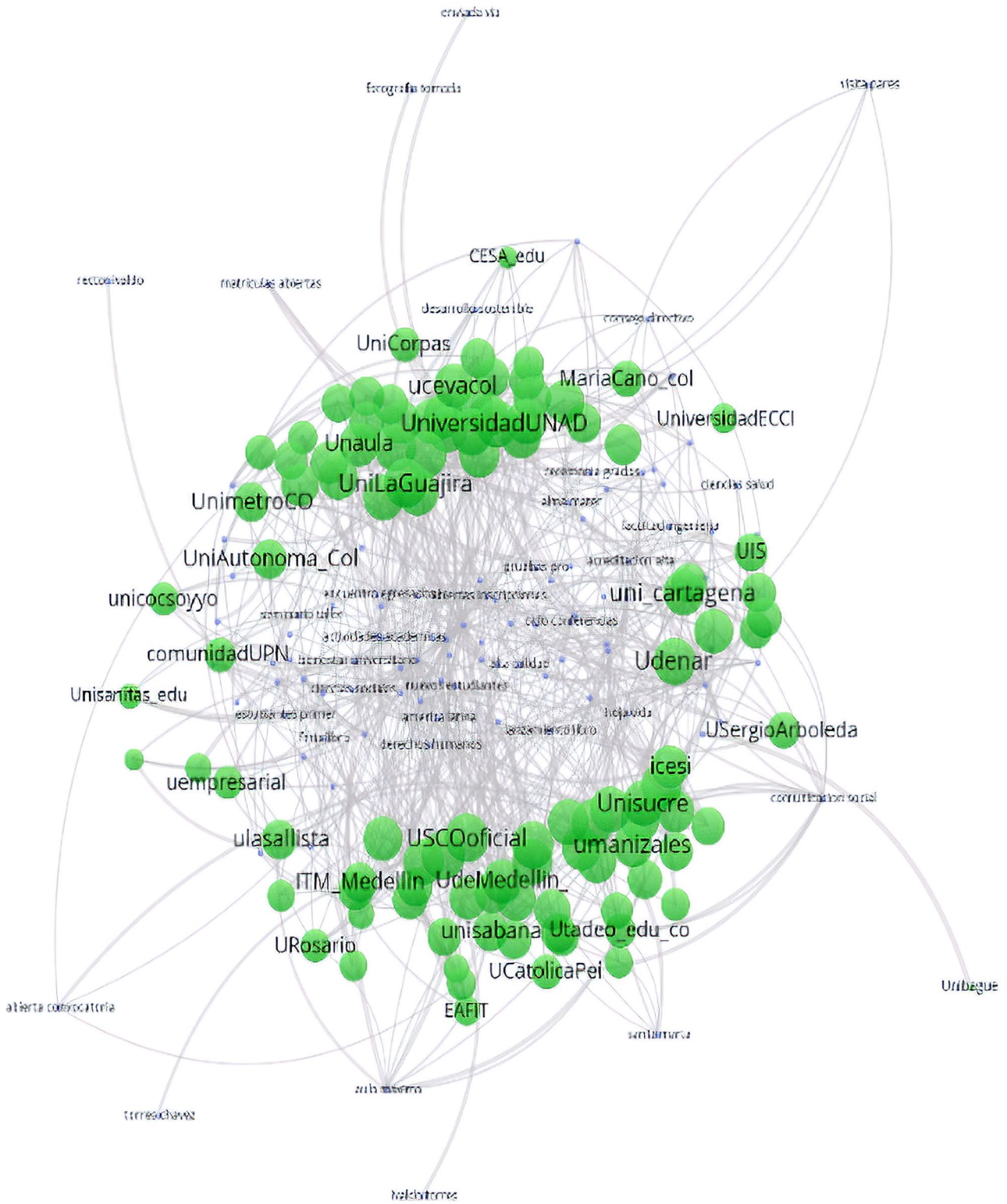

Fig. 10 Universities centrality

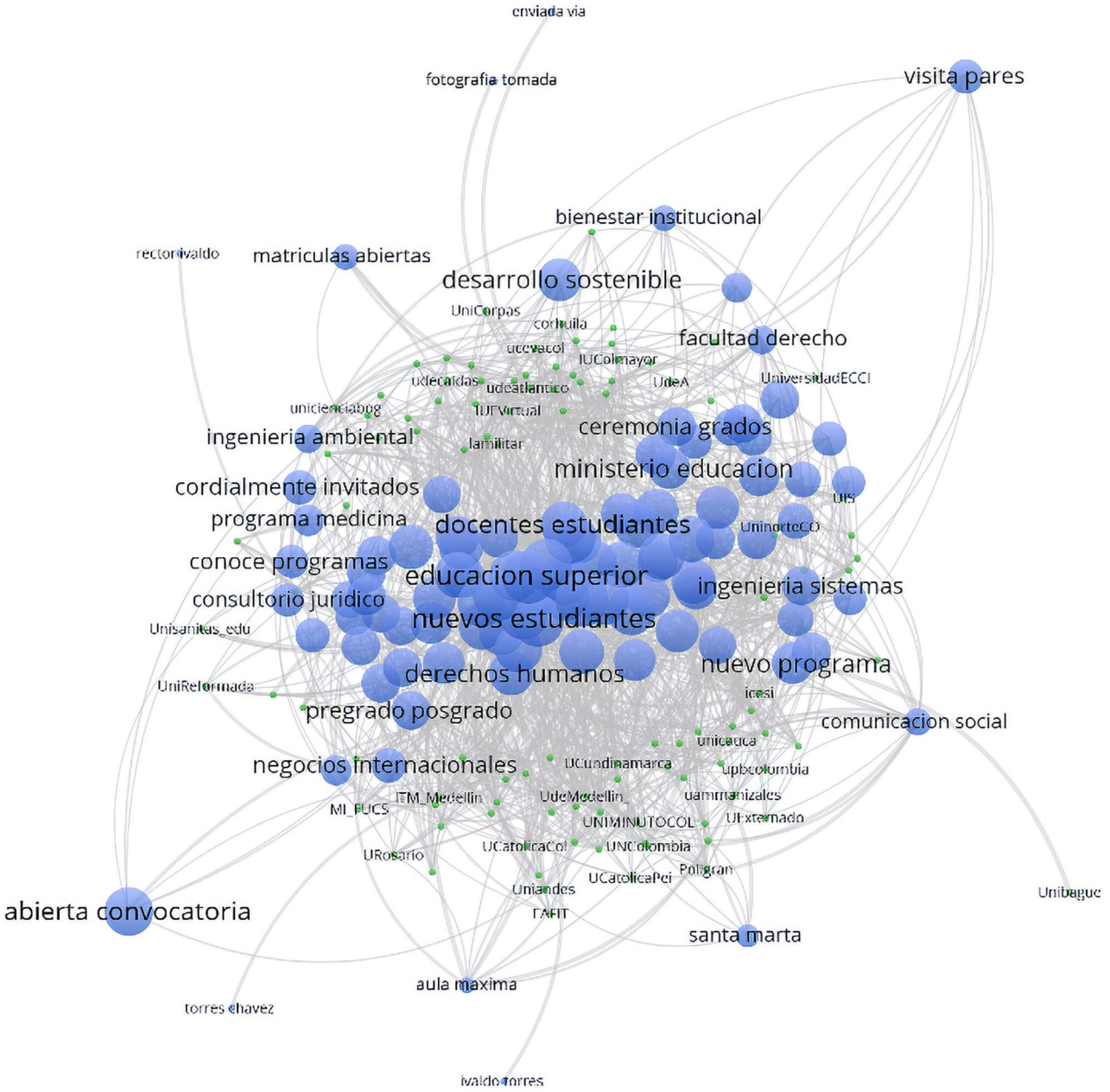

Fig. 11 Bigram centrality

Conclusions

A shared language is a key distinctive feature of a human community. The present work has reached good results through an approach to community detection in digital networks, inspired on the idea of a shared language rather than other kind of connections among community elements. Furthermore, a tailored study of a community's language allowed us to detect specific sociolinguistic features that might reveal relevant vocabulary which is not shared with other communities, even in close geographies or demographics. This vocabulary, in turn, can be attached to specific topics, motivations or interests of the community, as shown in our case study. Better understanding of specific communities and, particularly, of their shared language, might be of central importance to tailor messages for them, or to detect important trends, whether in public health, marketing, politics or national security. This work is a step forward in this direction, taking into account the difficulties of limited data in specific communities.

Some limitations should be pointed out. Shared language might be heavily influenced by topics, which makes

it difficult to differentiate community specifics from common interests. On the other hand, a broader community might go undetected due to subtle variations in vocabulary or language changes (for instance, from Spanish to English). Further research should cover these difficulties, including community detection work based on both links (such as mentions) and shared language comparisons.

Acknowledgements We would like to thank the Center for Excellence and Appropriation in Big Data and Data Analytics (CAOBA), Pontificia Universidad Javeriana, and the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC). The models and results presented in this challenge contributed to the building of the research capabilities of CAOBA. Also, the author Edwin Puertas gives thanks to the Universidad Tecnológica de Bolívar.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Dumbill E. A revolution that will transform how we live, work, and think: An interview with the authors of big data. Big data. 2013;1(2):73–7.
2. Meyerhoff M. Introducing sociolinguistics. Taylor & Francis Group: Routledge; 2015.
3. Meyerhoff M. Introducing sociolinguistics. Routledge; 2018.
4. Scott J. Social network analysis: developments, advances, and prospects. Social network analysis and mining. 2011;1(1):21–6.
5. Zeinab Kafi, Khalil Motallebzadeh. An introduction to sociolinguistics. International Journal of Society, Culture & Language. 2016;4(2):134–40.
6. Bryden J, Funk S, Jansen VA. Word usage mirrors community structure in the online social network twitter. EPJ Data Science, 2013;2(1):3.
7. Ríos SA, Muñoz R. Dark web portal overlapping community detection based on topic models. In Proceedings of the ACM SIGKDD workshop on intelligence and security informatics. 2012. p. 1–7.
8. Nguyen D. A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. Computational sociolinguistics: A survey Computational linguistics. 2016;42(3):537–93.
9. Reynolds WN, Salter WJ, Farber RM, Corley C, Dowling CP, Beeman WO, et al. Sociolect-based community detection. In 2013 IEEE International Conference on Intelligence and Security Informatics. 2013. p. 221–226, IEEE.
10. Mansouri F, Abdelalim S, Ikram EA. A modeling framework for the moroccan sociolect recognition used on the social media. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications. ACM. 2017. p. 34.
11. Gibson KR. Tool use, language and social behavior in relationship to information processing capacities. Tools, language and cognition in human evolution. 1993. p. 251-269.
12. K Adnan, R Akbar. An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data. 2019;6(1):91.
13. Louwerse MM. Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. Computers and the Humanities. 2004;38(2):207–21.
14. Paradis RD, Davenport D, Menaker D, Taylor SM. Detection of groups in non-structured data. Procedia Computer Science. 2012;12:412–7.
15. A Hussain, E Cambria. Semi-supervised learning for big social data analysis. Neurocomputing. 2018;275:1662–733.
16. Li L, Wu L, Evans JA. Social centralization and semantic collapse: Hyperbolic embeddings of networks and text. CoRR, abs/2001.09493, 2020.
17. Balaanand M, Karthikeyan N, Karthik S, Varatharajan R, Manogaran G, Sivaparthipan C. An enhanced graph-based semi-supervised learning algorithm to detect fake users on twitter. The Journal of Supercomputing. 2019;75(9):6085–105.
18. Cavallari S, Cambria E, Cai H, Chang KC, Zheng VW. Embedding both finite and infinite communities on graphs [application notes]. IEEE Computational Intelligence Magazine. 2019;14(3):39–50.
19. H Fani, E Jiang, E Bagheri, F Al-Obeidat, W Du, M Kargar. User community detection via embedding of social network structure and temporal content. Information Processing & Management. 2020;57(2):102056.
20. Park C, Han J, Yu H. Deep multiplex graph infomax: Attentive multiplex network embedding using global information. Knowledge-Based Systems. 2020. p.105861.
21. Liu P, Zhang L, Gulla JA. Real-time social recommendation based on graph embedding and temporal context. International Journal of Human-Computer Studies. 2019;121:58–72.
22. Tkachenko N, Guo W. Conflict detection in linguistically diverse on-line social networks: A russia-ukraine case study. In Proceedings of the 11th International Conference on Management of Digital EcoSystems, MEDES '19. Association for Computing Machinery. New York, NY, USA. 2019. p. 23-28.
23. E Cambria. Affective computing and sentiment analysis. IEEE intelligent systems. 2016;31(2):102–7.
24. Poria S, Chaturvedi I, Cambria E, Bisio F. Sentic lda: Improving on lda with semantic similarity for aspect-based sentiment analysis. In 2016 international joint conference on neural networks (IJCNN). 2016. p. 4465–4473, IEEE.
25. Hevner A, Chatterjee S. Design research in information systems: theory and practice. Springer Science & Business Media. 2010;2.
26. González RA, Pomares A. La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería. Reunión Nacional ACOFI. 2012. p. 12–14.
27. Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS. Social media? get serious! understanding the functional building blocks of social media. Business horizons. 2011;54(3):241–51.
28. Española RA. Banco de datos (CREA). Corpus de referencia del español actual. 2015. p. 2011–10.
29. Spitkovsky VI, Alshawi H, Chang AX, Jurafsky D. Unsupervised dependency parsing without gold part-of-speech tags. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. Edinburgh, Scotland, UK. 2011. p. 1281–1290.
30. Khurshid A, Gillam L, Tostevin L. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In The Eighth Text REtrieval Conference (TREC-8). Gaithersburg, Maryland. 1999. p. 1–8.
31. Joseph K, Carley KM, Hong JI. Check-ins in blau space applying blau macrosociological theory to foursquare check-ins from new york city. ACM Transactions on Intelligent Systems and Technology (TIST). 2014;5(3):1–22.

32. Park Y, Alam MH, Ryu WJ, and Sangkeun Lee. Bl-lda: Bringing bigram to supervised topic model. In 2015 International Conference on Computational Science and Computational Intelligence (CSCI). 2015. p. 83–88, IEEE.
33. Camacho D, Panizo-LLedot A, Bello-Orgaz G, Gonzalez-Pardo A, Cambria E. The four dimensions of social network analysis: An overview of research methods, applications, and software tools. Information Fusion. 2020;63:88–120.
34. Varelo AR. Hacia un modelo de aseguramiento de la calidad en la educación superior en colombia: estándares básicos y acreditación de excelencia. Educación superior, calidad y acreditación. CNA., 2003.
35. Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation. Machine learning. 1999;34(1–3):177–21010.
36. Damani OP, Ghonge S. Appropriately incorporating statistical significance in pmi. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2013. p. 163–169.
37. Arora S, Li Y, Liang Y, Ma T, Risteski A. A latent variable model approach to pmi-based word embeddings. Transactions of the Association for Computational Linguistics. 2016;4:385–99.
38. Ahmad K, Gillman L, Tostevin L. Weirdness indexing for logical document extrapolation and retrieval. In Proceedings of the Eighth Text Retrieval Conference (TREC-8). 2000. p. 1–8.