ISSN 2007-9737

Mining a Trending Topic: U.S. Immigration on the Context of Social Media

Esteban Castillo[1,*], Ofelia Cervantes[2]

[1] Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
Mexico

[2] Universidad de las Américas Puebla,
Department of Computer Science,
Mexico

esteban.castillojz@tec.mx, ofelia.cervantes@udlap.mx

Abstract. This paper presents a text mining approach for extracting valuable patterns from social media documents in the context of U.S. immigration. The paper points out the uncovering of statistical features alongside linguistic elements based on graph techniques. The use of graphs provide rich data structures for representing lexical and syntactic aspects of texts, allowing the discovery of complex patterns that used by experts could provide valuable insight. The proposed method is applied over a Twitter-X/-Reddit dataset that comprise English and Spanish language samples from 2016 up to 2019. Experimental results showed that our interpretation of classic statistic techniques provide a baseline understanding of the topic while a more robust analysis (graphs) permits to uncover/predict hidden patterns over large amount of samples. In particular, the use of a co-occurrence graph helped to obtain relevant words, phrases and sentences while a user-interaction graph allow to detect important users, communities and interactions among themselves.

Keywords. Text mining, statistics, graph mining, social network analysis, natural language processing, big data.

1 Introduction

Social media sites are an essential information resource related to every topic/domain around the world. Part of their success is due to the inherent openness for public consumption, clean and structured data, rich developer tooling, and broad appeal to users from every walk of life. Among the vast amount of available data on this sites, finding what is trending and how it is being discussed has emerged as an essential tool for understanding how people connect and how they share ideas, attitudes and even media consumption toward specific topics.

Text mining methods arrived as an optimal solution for acquiring, analyzing and predicting textual patterns from large amounts of data on social media. Mining methods, provide insightful knowledge that can be used by different domain experts for understanding user-profiles, authorship-styles, demographic information, sentiment polarity and even complex patterns related to the semantics of data.

Different studies around text mining have showed the impact of statistical and social network analysis for detecting insightful knowledge. Despite the progress achieved, there are still opportunities to create alternative approaches for representing and extracting complex patterns based on the combination of classic statistics and graph representations.

Considering the above, this paper proposes a text mining approach for identifying text patterns based on statistics and graph mining techniques over a highly commented topic (U.S. immigration).

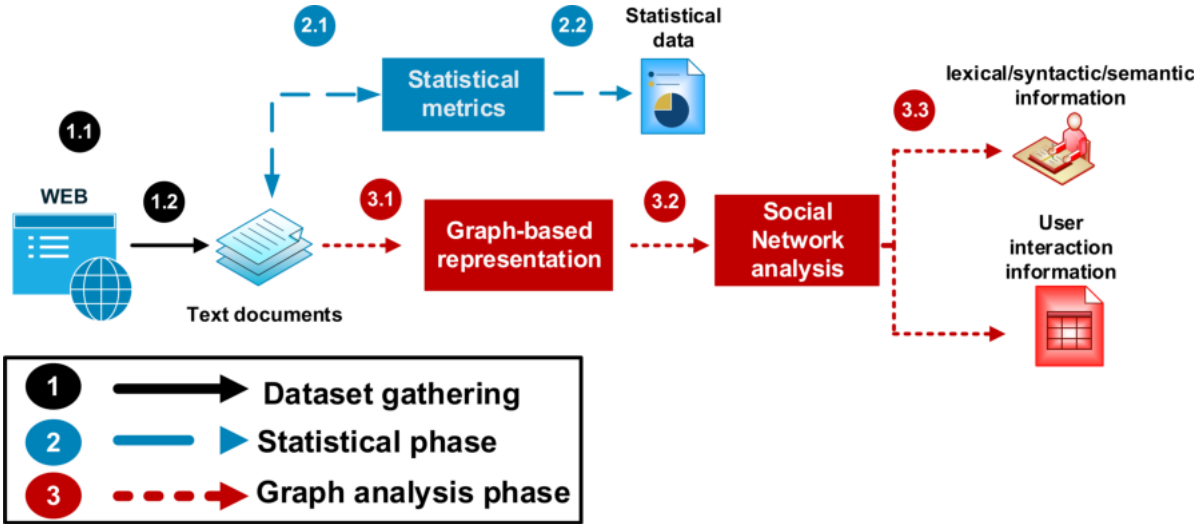

Fig. 1. Proposed text mining process

The **approach contribution** relies on the extraction/understanding of patterns and the creation of graph-based representations to detect knowledge using social network analysis tools. The **hypothesis** is that statistic techniques combined with more refined data structures (graphs) could be suitable for detecting representative elements over texts.

The **approach motivation** is to provide a valuable guide for mining a topic by using distinct text tools that normally are not combine together over a specific problem. The remainder of this paper is structured as follows:

Section 2 present existing approaches that deal with the extraction of knowledge from social media by using text mining techniques. Sections 3 to 6 provide details and examples on the design and implementation of the approach. Finally, implications and conclusions derived from this work thus far are presented in Section 7.

2 Related Work

Many literature deals with the extraction and digesting of social media on different topics and domains [3]. Most of it, is based on the use of classic statistic metrics, Information Retrieval (IR), Natural Language Processing (NLP), Machine learning (ML) and ultimately text mining methods. This literature, is highly dense and cover different applications and methodology approaches. Therefore, related work could be seen from two main perspectives: overall text mining approaches and specific research avenues related to social media on the context of political documents.

2.1 Text Mining

Text mining is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights in data [21]. The process of text mining involves the use of different research methods for obtaining valuable information from large amounts of data [27].

Some of this methods include techniques for acquiring and analyzing digital documents with NLP or statistics (besides others) [31, 5]. Researchers today, are using distinct text mining approaches for predicting domain patterns, public opinion and collectible behavior [13].

Text mining software has impacted the way that many industries work, allowing them to improve user experience and business decisions. Examples of this impact can be seen in different areas like customer service [23] where chatbots, and profiling tools are making the user experience

Table 1. Immigration keywords ordered by frequency of occurrence

English keywords obtained			
immigration$_1$	migration$_2$	naturalization$_3$	deportation$_4$
passport$_5$	greencard$_6$	border$_7$	trump$_8$
embassy$_9$	patrol$_{10}$	mexican$_{11}$	american$_{12}$
biden$_{13}$	workforce$_{14}$	alien$_{15}$	
Spanish keywords obtained			
inmigracion$_1$	migrante$_2$	indocumentado$_3$	repatriacion$_4$
deportacion$_5$	paisano$_6$	remesa$_7$	pasaporte$_8$
mexico$_9$	trump$_{10}$	frontera$_{11}$	epn$_{12}$
usa$_{13}$	migrante$_{14}$	amlo$_{15}$	

Table 2. Immigration dataset from July 11th of 2016 to July 11th of 2019

Social Media	**Language**	**Feature**	**Value**
X (Twitter)	English	Number of properties per tweet	71
		Number of tweets	4,412,621
		Number of geolocated tweets	1,755,468
		Avg. tweets per day	19,296.35
		Avg. geolocated tweets per day	8,471.6
		Avg. words per tweet	17.64
		Avg. file size per day	87.71 MB
	Spanish	Number of properties per tweet	71
		Number of tweets	1,974,944
		Number of geolocated tweets	732,835
		Avg. tweets per day	2,466.25
		Avg. geolocated tweets per day	1,522.67
		Avg. words per tweet	21.96
		Avg. file size per day	10.13 MB
	Number of tweets (both languages)		6,387,565
Reddit	English	Number of properties per post	15
		Number of posts	2,563,812
		Avg. posts per day	2501.98
		Avg. words per post	59.26
		Avg. file size per day	3.15 MB
Total number of texts about immigration			**8,951,377**

faster and simple; Healthcare systems [11], where distinct tools collect massive amounts of medical information for detecting critical insight of patients; and Cyber-security [15], where spam filtering and automatic intruder detection tools are making possible to identify malicious users. Text mining techniques are becoming more integrated and easier to use on the web.

This in turn, have introduce many approaches related to trending elements. Examples of this include: the analysis of digital marketing [39], recommendation systems [4], decision-making [36] and social network analysis [19]. For this last one, the extraction of non-trivial knowledge related to what is trending and whats is not have made an special effect on how users understand and consume information.

Considering its impact. it can be seen the importance of text mining and how it help others to make the most of their data, which leads to better decisions. Without this kind of tools, it would be impossible to analyze massive amounts of information (mainly on the web) which in consequence will stop the growing-flow of knowledge on the web.

2.2 Text Mining on Political Documents

In the context of social media (X, Facebook, Reddit, etc.), the analysis of political data by text mining techniques [22] have gained momentum considering the large amount of textual information, the number of interactions and the importance of the topic today (specifically on the U.S.) [29].

Citizenship and law enforcement topics have different text mining approaches that deal with the extraction of relevant users using social network measures (closeness, degree centrality, etc.) [31, 14]. Other approaches [32] used different ML algorithms (triplets and clustering) for detecting communities that might be important in a specific context. Other kind of approaches use classic statistic metrics like mean, median, mode, etc combined with IR scrapers for detecting structural patterns on social media data [18].

Immigration and border Security topics [7] also have distinct text mining approaches like the use of a friend of a friend (foaf) and co-occurrence graphs for obtaining activity patterns associated to users [17] or for summarizing/understanding large amounts of textual information [33]. Other kind of techniques rely on NLP for applying Part of Speech (PoS tags) and entity recognition techniques for scrapping linguistic features that help to understand the structure and purpose

Table 3. Immigration statistics: Baseline analysis

Statistic metric	Result	Description
Average number of words in X and Reddit	X: 16.81 Reddit: 87.62	Amount of words found in texts.
Average number of phrases in X and Reddit	X: 9.52 Reddit: 47.13	Groups of words that form meaningful units within a sentence.
Average number of sentences in X and Reddit	X: 5.52 Reddit: 26.13	Groups of words that make complete ideas.
Average Word Length	X: 6.34 Reddit: 4.76	Used to see how usual/unusual are the words in texts.
Number of different words on English language	X: 1,320,491 Reddit: 128,544	Distinct English words used in the immigration text documents.
Number of different words on Spanish language	X: 211,694 Reddit: 0	Distinct Spanish words used in the immigration text documents.
How users start a sentence	X: RT Reddit: What	X: Most of texts are a reply of other users. Reddit: Most of posts are questions related to immigration.
How users end a sentence	X: URL Reddit: ?	X: Most of texts finish with a reference to the source of the information. Reddit: Most of posts finish with the question mark, reinforcing the theory that most Reddit posts are questions
Most frequent PoS tags on X	Nouns Adjectives Verbs	References (nouns) to persons, places, things, or ideas are the most frequent words.
Most frequent PoS Tags on Reddit	Adjectives Nouns Adverbs	Words (adjectives) that describe or clarify nouns are the most frequent words.
Number of different users (@) on X	English: 68,447 Spanish: 13,671	Diversity of users that talk about immigration on X.
Number of different subtopics (#) on X	English: 8,590 Spanish: 1642	Diversity of subtopics related to immigration on X.
Number of different web sources (URLs) on X	English: 448,420 Spanish: 35,546	Diversity of URLs related to immigration on X.
Number of different web sources (URLs) on Reddit	12,674	Diversity of URLs related to immigration on Reddit.
Text documents distribution	X Eng: 79.7% X Spa: 10.2% Reddit Eng: 10.1%	Percentage of text documents in the dataset.
Geolocated Text documents distribution on X	English: 82.1% Spanish: 17.9%	Percentage of Geolocated documents in the dataset.

Table 4. Immigration statistics: Most/less frequent words

Frequent words on the dataset (X and Reddit)			
Ranking	**X English**	**X Spanish**	**Reddit English**
1	rt	rt	visa
2	immigration	pasaporte	greencard
3	trump	paisano	usa
4	passport	mojado	immigration
5	nafta	migrante	work
Unusual words on the dataset (X and Reddit)			
Ranking	**X English**	**X Spanish**	**Reddit English**
1	brown-skinned	vacaciones	season
2	neighbourhood	tramite	paperwork
3	muslim	preguntas	article
4	traveling	tramite	reasons
5	Rusia	ilegalidad	telephone
Most frequent users and topics on X			
Ranking	**X users (@)**	**X topics (#)**	
1	realDonaldTrump	immigration	
2	BreitbartNews	Trump	
3	HillaryClinton	MAGA	
4	FoxNews	migration	
5	FAIRImmigration	ny	
Less frequent users and topics on X			
Ranking	**X users (@)**	**X topics (#)**	
1	SarahPinder2	CNNIdiots	
2	tatianashanks	RefugeeRights	
3	ErfanSoomro	Illuminati	
4	HoopsmanB	yeahRight	
5	RamblinGrimace	FeelingSad	

of information (attitude, entities, sentiments etc.) [20]. Other approaches [6] have deal with the immigration analysis by detecting geo-spatial patterns of users on the Mexico-U.S. border. Additionally, distinct efforts [2] have implemented mining techniques for automatic content-characterization of news, stories and blogs related to the immigration phenomena.

From the different mining approaches implemented for political purposes, it can be seen that is a growing area where more interactions are available everyday. This in turn have made possible to analyze relevant text patterns from different social media sources, which applied in different research and decision-making process have a meaningful impact on how users understand, digest and use knowledge.

3 Text Mining Process

Data mining [41] and text mining are similar in terms of the way they extract valuable insight from data. The first one focus on the analysis of distinct data types (texts, images, sound, etc.) while the second one focuses only on the retrieval of textual information. Despite this key difference, both mining approaches have similar steps involved in the knowledge discovery process with the exception that in text mining, special emphasis is made on the data modeling considering the unstructured nature of texts and the different linguistic aspects to explore. Taking that in mind, Figure 1 shows the proposed steps to retrieve text patterns on the context of the U.S. immigration phenomena. The approach consists of three overall steps:

1. Dataset creation (see Section 4)
   - 1.1 Text acquisition: Create a large collection of text documents using different social media resources: X, Reddit, etc.
   - 1.2 Text preprocessing: Preprocess documents to guarantee homogeneity among texts.
2. Statistical analysis: (see Section 5)
   - 2.1 Choose distinct statistical metrics depending of the specifics of text documents.
   - 2.2 Extract statistical information of texts that describes and explores the nature of the data on the underlying topic.
3. Graph analysis: (see Section 6)
   - 3.1 Create rich representations for the text documents (see Sections 6.1 and 6.2).
   - 3.2 Uncover insightful knowledge from data representations using social network analysis tools.

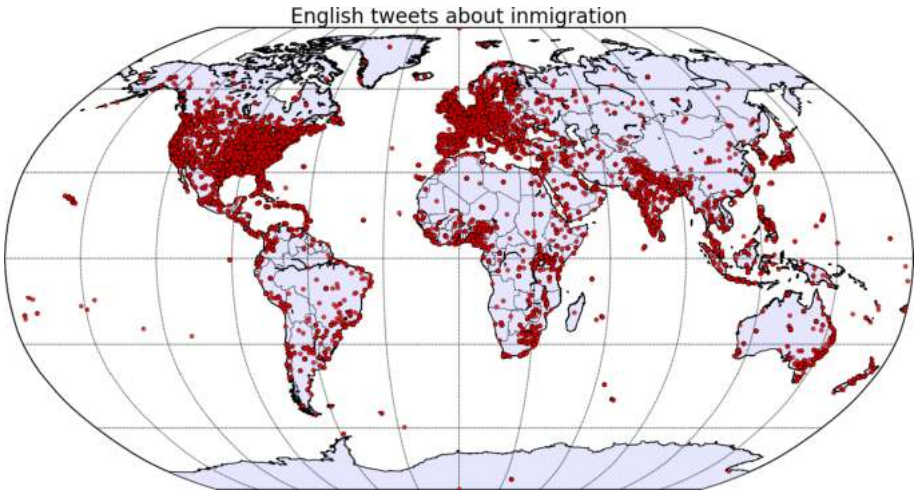

Fig. 2. Immigration statistics: English language texts distribution on X

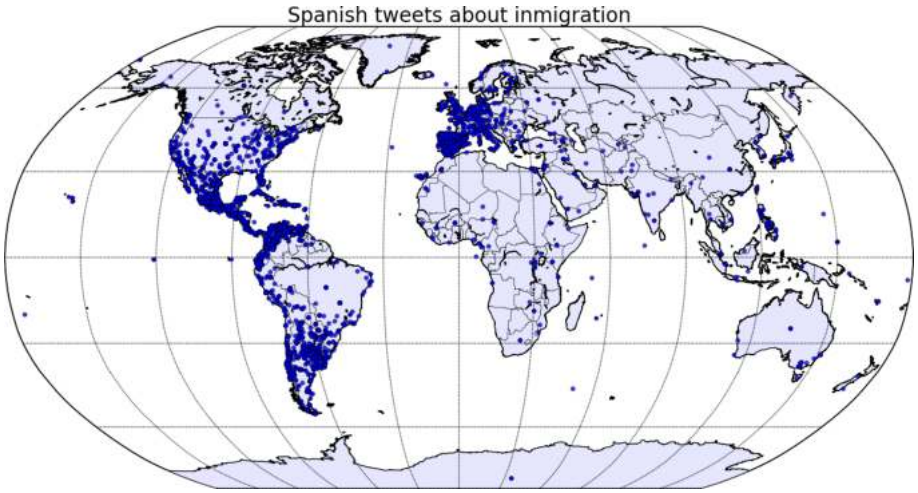

Fig. 3. Immigration statistics: Spanish language texts distribution on X

3.3 Use the obtained information to characterized in a better way a trending topic (see Sections 6.4 and 6.5).

From the previous figure, the steps associated to the analysis of the immigration phenomena are presented. The first step deals with the information gathering from distinct social media channels, making special emphasis on text acquisition and cleaning. The second step deals with the empirical analysis and inference of knowledge by using classic statistical metrics.

Finally, the third step involves the use of robust data structures (graphs) and social network analysis tools to uncover relevant patterns that statistics are unable to discover.

4 Dataset Creation

In this section, the collection of text documents associated to the immigration topic on English and Spanish languages is discussed. First, the keywords used for extracting text samples from

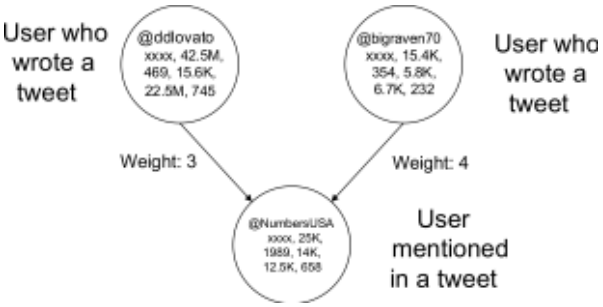

Fig. 4. User interaction graph example

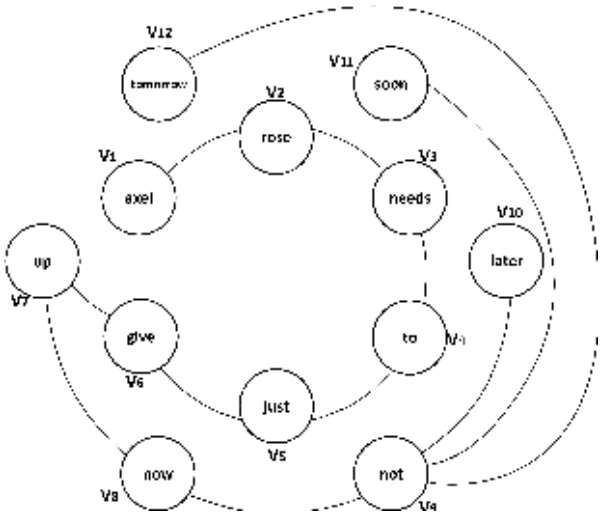

Fig. 5. Example of a co-occurrence graph with a window of two words

social media through different APIs (Application Programming Interfaces) are shown. Later, the description of the chosen social media sources used and the type of text documents obtained are presented. Finally, the main dataset features associated to the immigration topic are displayed considering the proposed text mining method (see Section 3).

4.1 Keyword Selection

In order to obtain suitable information from social media, words related to the immigration topic on July 2016 (dataset starting point) were extracted according to their frequency of occurrence on one thousand web pages[1]. For each language, the fifteen keywords without stopwords[2] or special characters are extracted from the web pages. These words were used as input for diverse social media APIs to obtain relevant text documents. Table 1 summarize words used on each language.

4.2 Social Media Sources

There are several social media APIs to obtain textual information, but most of them do not offer a public streaming to download texts periodically. Among the media channels that do not have this kind of restriction, X (formerly known as Twitter)[3] and Reddit[4] provide robust tools to download streaming data with minimum authentication and authorization from users over different languages.

Additionally, both media channels have real time interaction among users and cover a vast amount of domains around the world (U.S. specifically). Considering the previous features and the free availability of data on both social media channels, it was decided to use X and Reddit as primary data sources for detecting valuable knowledge related to the Immigration topic.

4.3 Text Documents Obtained

Using X and Reddit APIs, a dataset that comprise English and Spanish samples from U.S. users were collected. The dataset contains text documents from July 11 2016 to July 11 2019. All dataset samples were collected in a daily basis using a JSON format[5] with a UTF-8 encoding for storage each text document. Each collected sample contains properties/metadata related to the textual interaction. In the case of X[6], seventy one properties are obtained for each sample including: name, country, date, number of followers, number of likes, etc.

[1] The webpage number was decided based on the Google ranking and the amount of textual information.
[2] Stopwords represents a group of words that bear no content or relevant semantics in the text.
[3] twitter.com/
[4] www.reddit.com/
[5] JSON is a syntax for storing and exchanging data o the web.
[6] dev.twitter.com/overview/api/tweets

Table 5. Immigration graphs: main properties

Social Media	Graph Type	Vertex Type	Number of Vertices	Edge Type	Number of Edges
X and Reddit in English language	Co-occurrence	Words	1,380,615	Two words appear together in the text.	690,843
X and Reddit in Spanish language	Co-occurrence	Words	40,386	Two words appear together in the text.	90,128
X in English language	User interaction	Users	97,582	One user reference other in the text.	135,295

For Reddit[7], fifteen properties are collected including: author name, text date, country, topic name, topic score, subreddit name, etc. The dataset also provides some geolocation properties. In the case of X, this provides coordinates and place elements (if these are release by the user) while for Reddit, this does not provide any kind of metadata that can be used for obtaining the latitude and longitude associated to a post due to some API restrictions.

Table 2 summarize the dataset main features emphasizing the number of documents for Spanish and English languages on X and Reddit. From the dataset table, it can be observed that the English subset considered both social media channels while for Spanish it is only used X.

This is due to the lack of Spanish samples and the small amount of texts retrieve from the Reddit API. Additionally, The number of samples collected highlight the amount of interactions in both social media channels. In the case of X, there are more small interactions (140 characters at most) while for Reddit there are less interactions but these condensate more textual information.

5 Statistical Analysis

As a first attempt to obtain valuable knowledge from text documents, a statistical analysis [35] was applied according to the proposed approach (see Section 3).

The main goal of this phase, was to find and summarize the presence of textual patterns to describe or estimate information that can be helpful to understand the nature of the topic in the context of social media. Table 3 display some baseline statistics found in both media channels. For each entry, the metric used, the result obtained and a brief description is presented. From the previous table, it can be noticed that baseline statistics provide valuable insight about the data collected.

The average number of words, phrases and sentences indicate how diverse is the vocabulary as well as the amount of ideas/sentences used to talk about immigration. How users start or end a text also illustrate the kind of interactions associated to the topic (is a question or a reply from other users). Additionally, The analysis of PoS tags[8] permits to understand about whats people focus more on textual interactions (entities-ideas or descriptive elements about the topic).

Other statistical elements can be seen in Table 4 where frequent and unusual words from X and Reddit are shown. The table display frequent users (@) and topics (#) obtained from X interactions, this information is useful for understanding what is trending and what is not considering the frequency of occurrence of immigration words. In the case of geolocated texts on X, Figures 2 and 3 show a statistical tweet distribution on English and Spanish languages.

[7] www.reddit.com/dev/api/

[8] POS tagging is the process of marking up a word based on the syntactic role that it plays in a sentence.

Table 6. User interaction graph: Top X users on English and Spanish

Centrality Measure	English X users (@)		
Degree	$RealDonaldTrump_1$	CNN_2	$FoxNews_3$
	$HillaryClinton_4$	$YouTube_5$	$BarackObama_6$
Closeness	$RealDonaldTrump_1$	$FoxNews_2$	CNN_3
	$Nytimes_4$	$CBCNews_5$	$NumbersUSA_6$
Betweenness	$RealDonaldTrump_1$	$HillaryClinton_2$	$CBCNew_3$
	$MSNBC_4$	$YouTube_5$	$Reuters_6$
Centrality Measure	**Spanish X users (@)**		
Degree	EPN_1	$lopezobrado_2$	$GenPenaloza_3$
	$lopezdoriga_4$	$SREmx_5$	$YouTube_6$
Closeness	$RedsocialSAIME_1$	$LeonKrauze_2$	SRE_3
	$ElNacionalWeb_4$	$elpais_5$	$YouTube_6$
Betweenness	$YouTube_1$	$CNNEE_2$	$Telemundo_3$
	$Univision_4$	$TwitterEspanol_5$	$ElUniversal_6$

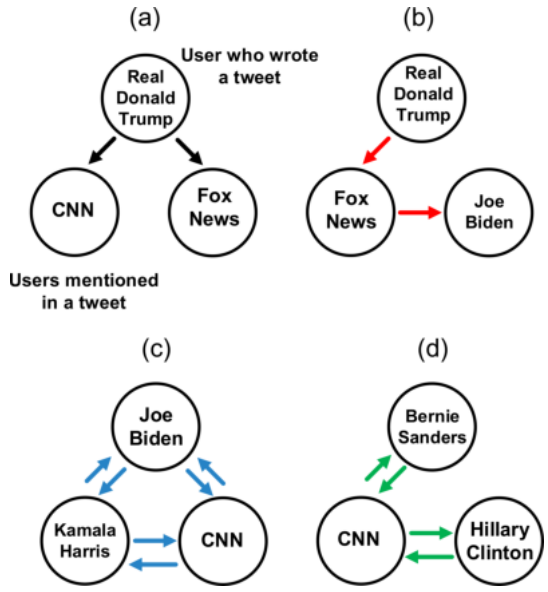

Fig. 6. User interaction graph: English triad examples found

In these figures, it can be observed that in the case of the English tweets most of U.S. users that talk about immigration were located (at time of posting) in North America, Europe and India. For the Spanish tweet distribution, the U.S. users that talk more about immigration were located in America (North America, Central America and South America) and Europe.

Taking into account some examples of statistics obtained from the dataset, it can be remarked the importance of this kind of elements for a basic understanding of the immigration phenomena. The analysis of frequent elements on X and Reddit provide an intuitive way for detecting relevant words, users, topics, etc. This in turn, provide insightful knowledge for implementing more advance data representations (graphs).

6 Graph Analysis

In this section are described the patterns obtained using more complex structures like graphs in the context of a text mining approach (see Section 3). First, two graph representations are proposed: one based on the interaction of words related to immigration (word co-occurrence) and another based on the interaction of users that talk about the topic. Latter, the patterns/subgraphs obtained are presented and described.

6.1 User Interaction Graph

Taking into consideration how users relate to each other and the importance of this relationships to understand social media synergy, a graph that represents the interaction among users that talk about immigration topics on X is proposed (Reddit does not give much importance to users as X does). Formally proposed graph $G = (V, E, L_V, L_E)$ has the following attributes:

1. $V = \{v_1, ..., v_n\}$ is a finite set of vertices that consists of the X users (@) contained in one or several texts.
2. $E \subseteq V \times V$ is the finite set of edges which represent that a user referenced other user or that other user referenced him in another text.
3. L_V is the label set of V, where each vertex has the following properties:
   - X user name: Name of the X user (@).
   - X user ID: Unique identifier for this text.
   - Followers count: The number of followers this account currently has.

Fig. 7. User interaction graph: Community of Spanish users found

- Friends count: The number of users this account is following (AKA their "followings").
- Statuses count: The number of text (including re-texts) issued by the user.
- Favorites count: Indicates approximately how many times a text has been "liked" by X users.
- Listed count: The number of public lists that this user is a member of.

4. L_E is the label/weight set of E, where:
L_E = #hashtags + #URLs + #interactions
that have in common two X users (two vertices).

As an example of this graph-based representation, consider the following texts extracted from two texts from X:

- **User1**: @ddlovato, **Text1**: @NumbersUSA #ElectionDay #IVotedBecause I believe in equality, we need comprehensive immigration.
- **User2**: @bigraven70, **Text2**: @NumbersUSA I'm all for legal immigration #immigration. My family came here legally and yours can too #Trump URL.

Based on proposed properties, user and text information previously preprocessed[9] can be mapped to the user interaction graph shown in Figure 4. In the user interaction graph, the communication (edges) between X users (vertices) that talk about the immigration in English is considered.

The goal of this graph, is to take advantage of the properties provided by X (and not by Reddit) like the users (@), topics (#) and even the web resources (URLs) to propose a weighting scheme that could be used to uncover meaningful users, communities and new interactions. This graph can be seen as a special foaf graph[10] [38] where the relationships are oriented to model how users relate to each other when they are mentioned in a specific context.

6.2 Co-Occurrence Graph

Keeping in mind the relevance of lexical elements (based on previous section) and the lack of syntactic structure on social media documents. A non-directed and unweighted graph representation based on the co-occurrence [31, 40] of two words is proposed.

The objective of this graph is to take advantage of all natural interactions between words[11] to extract valuable lexical-syntactical patterns that can not be obtain using traditional statistics. Formally, the proposed co-occurrence graph used in the experiments is represented by $G = (V, E, L_V)$,where:

1. $V = \{v_1, ..., v_n\}$ is a finite set of vertices that consists of the words contained in many texts.
2. $E \subseteq V \times V$ is the finite set of edges which represent that two vertices are connected if their corresponding lexical units co-occur within a window of two words in the text at least once.
3. L_V is the label set of V, where $L_V = \{\text{etq} : \text{etq} \in \text{words}\}$

As an example, consider the following sentence ζ extracted from a text T: "Axel Rose needs to just give up. Now. Not later, not soon, not tomorrow.",

[9] Something similar to the co-occurrence graph.

[10] Friend of a friend graph structure.

[11] The bond of one term over another one in the syntactic order.

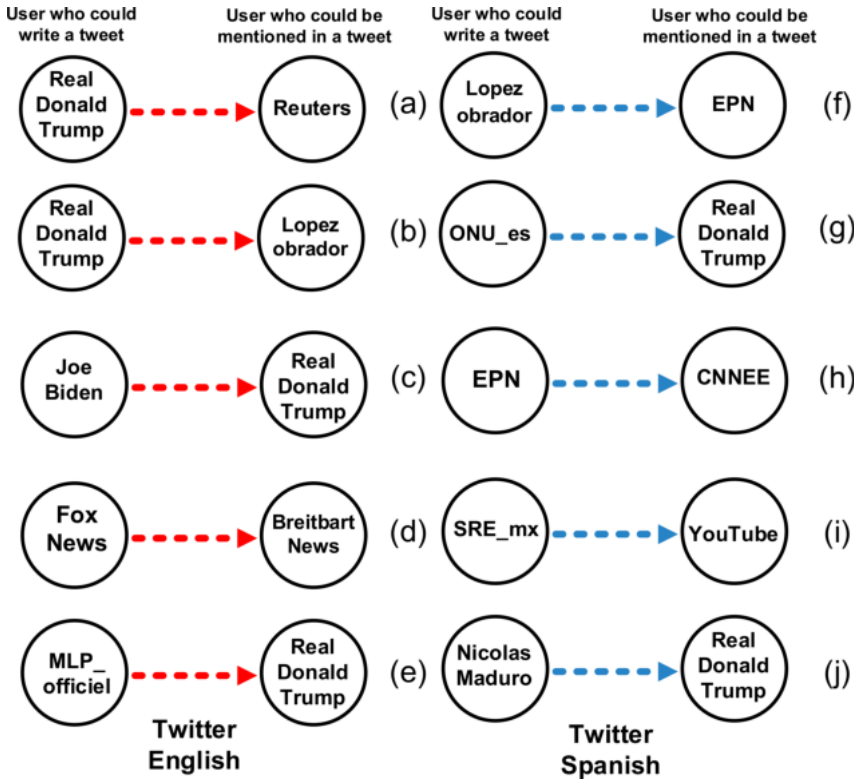

Fig. 8. User interaction graph: Link prediction examples found

which after the preprocessing stage[12] would be as follows: “axel rose needs to just give up now not later not soon not tomorrow”. Based on the proposed representation, preprocessed sentence ζ can be mapped to the co-occurrence graph shown in Figure 5.

The proposed co-occurrence graph captures the syntactic interactions (edges) of all the words (vertices) used when people discussed about immigration in X-Reddit (for English and Spanish languages). The idea of this graph is to obtain relevant words, phrases and even sentences that describe the way individuals write about the topic for understanding attitudes, trends and media consumption.

[12]This task includes lowercase all words in the texts and elimination of punctuation symbols that are not part of the ASCII encoding (except for @, # and URLs in the case of X).

6.3 Graph Properties

In order to understand the richness of the two proposed graph structures (see Sections 6.1 and 6.2). Table 5 shows main features associated to the co-occurrence graph and the user interaction graph in the context of the social media sites analysed. From table 5, it can be observed the amount of vertices and edges created from the social media dataset. This highlight the diversity of information obtained trough several months as well as the complexity of the data structures created. Considering graph properties, upcoming sections show results examples obtained by optimized graph mining algorithms [30].

6.4 User Interaction Graph Results

As in the statistical phase, the main objective of the user interaction graph is to obtain insightful knowledge from social media.

Table 7. Co-occurrence graph: Top words on X and Reddit

Centrality Measure	**English Language Words**			
Degree	RT_1	$immigration_2$	$Trump_3$	$passport_4$
	USA_5	$migration_6$	$people_7$	$visa_8$
Closeness	$immigration_1$	$Trump_2$	$election_3$	$Clinton_4$
	$Illegal_5$	$passport_6$	$Obama_7$	$Jobs_8$
Centrality Measure	**Spanish Language Words**			
Degree	RT_1	$pasaporte_2$	$paisano_3$	$indocumentado_4$
	$migrante_5$	$EEUU_6$	$frontera_7$	$Europa_8$
Closeness	$visa_1$	$trabajo_2$	$drogas_3$	$migrante_4$
	$pasaporte_5$	$frontera_6$	$remesa_7$	$Trump_8$

In this case, there are extract relevant users from X based on distinct social network analysis metrics (centrality, community detection and link prediction). Table 6 show top X users in English and Spanish languages using a reinterpretation of a classic social network analysis called centrality measures[13] [9]:

- **Degree centrality**: Collect users who are more referenced on immigration texts and users who referenced many others about the topic.
- **Closeness centrality**: Obtain users who are referenced immediately about immigration and users who spread out information faster about the topic.
- **Betweenness centrality**: Select people who have access to a lot of users that talk about immigration.

Other elements retrieved from the user interaction graph are triad elements [24]. This provide topological insight of the interaction of three interconnected vertices. These kind of subgraphs helps to understand how users spread information about immigration and how close they are in terms of their interactions.

[13] Graph Centrality refers to a family of structural measures related to the position/importance of vertex in a graph representation.

The triads with highest edge weight are consider as relevant communities that maintain a constant communication flow. Figure 6 display some examples of triads found on English language (something similar is performed for Spanish language). In addition to the triad detection, the uncovering of bigger communities is also performed by the user interaction graph.

The edge-betweenness centrality [37] is used for detecting high interconnected vertices on the network. The idea of this technique, is to gradually remove the weighted edges with highest betweenness and recalculate the centrality for all edges after every removal. This way sooner or later the network falls off into smaller components which are relevant user groups that talk about immigration.

Figure 7 illustrates the types of user communities obtained on the context of the immigration topic for the Spanish language (something similar is performed for English language). Finally, there is a applied an link prediction approach [28, 8] for inferring new relationships among users based on the topological structure of the graph and the weighted scheme proposed.

This kind of technique analyzed the neighborhood of vertices on the graph. Vertices that have similar neighbors will have new edges in the near future while vertices that do not share

Table 8. Co-occurrence graph: Top phrases on X and Reddit

Centrality Measure	**English Language Words**			
Degree	RT_1	$immigration_2$	$Trump_3$	$passport_4$
	USA_5	$migration_6$	$people_7$	$visa_8$
Closeness	$immigration_1$	$Trump_2$	$election_3$	$Clinton_4$
	$Illegal_5$	$passport_6$	$Obama_7$	$Jobs_8$
Centrality Measure	**Spanish Language Words**			
Degree	RT_1	$pasaporte_2$	$paisano_3$	$indocumentado_4$
	$migrante_5$	$EEUU_6$	$frontera_7$	$Europa_8$
Closeness	$visa_1$	$trabajo_2$	$drogas_3$	$migrante_4$
	$pasaporte_5$	$frontera_6$	$remesa_7$	$Trump_8$

many neighbors will remain without change. Figure 8 show some examples of interactions inferred using the link prediction and the user interaction graph on the English language.

6.5 Co-Occurrence Graph Results

The use of a co-occurrence graph permit to obtain insightful knowledge from texts that comes from different social media. The co-occurrence representation allow to extract more complex patterns that do not depend entirely of the frequency of occurrence of texts (like in statistics). Tables 7, 8 and 9 show relevant examples of words, phrases and sentences obtained by using the co-occurrence representation and some of centrality measures[14]. The following interpretations are proposed for extracting elements from the graph:

- **Words**: Top ranked words by degree and closeness centralities are extracted considering that such elements could be highly mentioned in the syntactic structure of immigration texts.
- **Small phrases**: Words that have a high interaction with other words, regardless of their syntactic relevance in the texts are suitable candidates to obtain collocations[15].

So, the top ranked vertices/words according to the degree centrality that are part of a collocation are obtained.

- **Sentences**: Considering that vertices with a high closeness centrality are words with an important role in the syntactic sequence of texts (they are reachable in the minimum number of steps), the sentences that have most of top ranked words according to this centrality are extracted.

In table 7, top words are presented without considering stopwords and special symbols. From the words obtained, it can be observed that degree centrality obtained words that people mentioned immediately when the immigration topic is discussed, while in the closeness centrality, the words that play an active and central role in the texts related to immigration are extracted.

In the case of the top phrases related to immigration, Table 8 shows the top collocations with stopwords but not special symbols. Analyzing the collocations obtained using the degree centrality it is possible to see the relevant subtopics used when individuals write about immigration on X and Reddit.

For top sentences related to immigration, Table 9 show some examples that reflect the most important attitudes and ideas related to immigration. These examples are formed of words with high closeness elements which implies that these sentences could be used as a brief summary of what people write about immigration.

[14] Among different centrality elements, degree and closeness demonstrate to be stable in the previous section.

[15] Pairs of words that always appear together in the texts.

Table 9. Co-occurrence graph: Top sentences on X and Reddit

Centrality Measure	English Language Sentences
Closeness	1. Your vote will make a difference. #vote
	2. I voted for Trump because illegal immigration is Illegal
	3. The Era Of Climate Migration Meets Violent Borders
	4. We are more concerned about immigration than any other nation
Centrality Measure	**Spanish Language Sentences**
Closeness	1. Presidencia de Trump despierta temores entre indocumentados
	2. No podemos esperar nada bueno de biden y kamala.
	3. ¿SERA? Mexico, preparado ante deportación de mexicanos
	4. Cancelación de la #deportacion de #EstadosUnidos?

7 Conclusions and Future Work

An approach that implements a text mining method based on statistical and graph analysis has been presented. Results obtained highlight the relevance of the implemented method and the importance of extracted patterns (statistical and graph ones) for explaining the nature of the immigration topic. Considering the theoretical implications of this mining method, the practical benefits associated are the following:

– The analysis of classic statistical metrics (frequency mainly) provide initial insight associated to immigration topic, and supply information for creating the graph-based representations proposed (like which centrality measure used considering the frequencies found).
– For mining the immigration topic, the user interaction graph and the co-occurrence graph with a window of two words show to be a very effective option to extract important information of texts, revealing that the co-occurrence graph not only works for classic NLP problems [10] but also for extracting distinct types of lexical/syntactical patterns on social media.

In the case of the user interaction graph, this showed to be a really good option to understand the dynamics of users in a specific network but it is necessary to test this kind of graph in other trending topics.

– The use of co-occurrence windows of two words allows to map the natural relationship of terms, which facilitate the analysis of lexical and syntactical elements of texts related to immigration.
– The co-occurrence graph can be used to extract topologically important words, phrases and even complete sentences that represent what people think or express when they write about the immigration topic.
– The user interaction graph permits to map the way in which X users relate to each other in the context of a specific topic, instead of just mapping the static relations that users have with others on a social network like in the classic friend of a friend graph.
– A weighted scheme on the interaction graph permits to evaluate in a more accurate way the user communication, considering the topics in common, the URLs and the frequency of occurrence between vertices.

This schema ultimately leads some graph mining algorithms (triad and community detection as well as link prediction) to find strong relationships between users that have multiple elements in common when they write about immigration.

– One of the major differences between the co-occurrence graph and the user interaction graph is that the co-occurrence of words can be applied to any kind of text document as long as the texts have a known encoding (like UTF-8 or ASCII). So, co-occurrence graphs offer more flexibility because they can be used in any text in any language with minimum preprocessing while the interaction graph permit an accurate analysis of X due to the social media specific properties.

Research on the use of a text mining approach continues in favor of improving obtained findings, keeping in mind the complexity of the use of graphs. Ongoing and future work includes the following actions:

– Work with experts on the U.S immigration problem to identify the impact of extracted patterns over decision making.
– Extract new features from graphs associated to the distinct levels of language to improve previous results [34].
– Experiment with other graph-based representations for documents that include semantic information related to texts [26].
– Applying different visualization methods on graph structures to present and understand obtained textual information in a more natural and easy-to-understand manner [25, 16].
– Analyze other trending topics like healthcare, news diffusion, etc. [1, 12], for testing the behavior of proposed graphs when applied to other real-world text documents.

Acknowledgments

This work has been partially supported by the CONACYT grant with reference #373269/244898. The authors would also like to thank Darnes Vilariño Ayala and David Báez López for their invaluable help reviewing this manuscript.

References

1. **Ahn, S. J., Yoon, H. Y., Lee, Y. J. (2021).** Text mining as a tool for real-time technology assessment: Application to the cross-national comparative study on artificial organ technology. Technology in Society, Vol. 66, pp. 101659. DOI: 10.1016/j.techsoc.2021.101659.
2. **Altarrazi, S. M., Sasi, S. (2016).** Tweeple's microblogs on illegal immigration in USA. International Conference on Electrical, Electronics, and Optimization Techniques, pp. 2011–2018. DOI: 10.1109/iceeot.2016.7755041.
3. **Balaji, T. K., Rao-Annavarapu, C. S., Bablani, A. (2021).** Machine learning algorithms for social media analysis: A survey. Computer Science Review, Vol. 40, pp. 100395. DOI: 10.1016/j.cosrev.2021.100395.
4. **Betancourt, Y., Ilarri, S. (2020).** Use of text mining techniques for recommender systems. Proceedings of the 22nd International Conference on Enterprise Information Systems, pp. 780–787. DOI: 10.5220/0009576507800787.
5. **Biemann, C., Mehler, A. (2014).** Text mining: From ontology learning to automated text processing applications. Springer International Publishing. DOI: 10.1007/978-3-319-12655-5.
6. **Borruso, G. (2009).** Geographical analysis of foreign immigration and spatial patterns in urban areas: Density estimation and spatial segregation. Lecture Notes in Computer Science, Vol. 5072, pp. 459–474. DOI: 10.1007/978-3-540-69839-5_34.
7. **Cartwright, K., Chacon, L. (2021).** The impact of immigration-related separation and reunification on children's education: Evidence from the american community survey 2010–2018. Children and Youth

Services Review, Vol. 126, pp. 106013. DOI: 10.1016/j.childyouth.2021.106013.

8. **Castillo, E., Cervantes, O., Vilariño, D. (2018).** Author profiling using a graph enrichment approach. Journal of Intelligent and Fuzzy Systems, Vol. 34, No. 5, pp. 3003–3014. DOI: 10.3233/jifs-169485.
9. **Castillo, E., Cervantes, O., Vilariño, D. (2019).** Authorship verification using a graph knowledge discovery approach. Journal of Intelligent and Fuzzy Systems, Vol. 36, No. 6, pp. 6075–6087. DOI: 10.3233/jifs-181934.
10. **Castillo-Juarez, E., Cervantes-Villagómez, O., Vilariño-Ayala, D. (2018).** Text analysis using different graph-based representations. Computación y Sistemas, Vol. 21, No. 4. DOI: 10.13053/cys-21-4-2551.
11. **Chatterjee, S., Goyal, D., Prakash, A., Sharma, J. (2021).** Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. Journal of Business Research, Vol. 131, pp. 815–825. DOI: 10.1016/j.jbusres.2020.10.043.
12. **Chen, W. K., Chen, L. S., Pan, Y. T. (2021).** A text mining-based framework to discover the important factors in text reviews for predicting the views of live streaming. Applied Soft Computing, Vol. 111, pp. 107704. DOI: 10.1016/j.asoc.2021.107704.
13. **Coenen, F., Fred, A., Aveiro, D., Dietz, J., Bernardino, J., Masciari, E., Filipe, J. (2023).** Knowledge discovery, knowledge engineering and knowledge management. 14th International Joint Conference. Springer Cham. DOI: 10.1007/978-3-031-43471-6.
14. **Cook, D. J., Holder, L. B. (2006).** Mining graph data. John Wiley and Sons.
15. **de-Boer, M. H. T., Bakker, B. J., Boertjes, E., Wilmer, M., Raaijmakers, S., van-der-Kleij, R. (2019).** Text mining in cybersecurity: Exploring threats and opportunities. Multimodal Technologies and Interaction, Vol. 3, No. 3, pp. 62. DOI: 10.3390/mti3030062.
16. **Evergreen, S. (2016).** Effective data visualization: The right chart for the right data. SAGE Publications.
17. **Fotouhi, B., Rabbat, M. G. (2012).** Migration in a small world: A network approach to modeling immigration processes. Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing, pp. 136–143. DOI: 10.1109/allerton.2012.6483210.
18. **Freire-Vidal, Y., Graells-Garrido, E. (2019).** Characterization of local attitudes toward immigration using social media. Companion Proceedings of The 2019 World Wide Web Conference, pp. 783–790. DOI: 10.1145/3308560.3316455.
19. **Gao, W., Sebastiani, F. (2015).** Tweet sentiment: From classification to quantification. Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 97–104. DOI: 10.1145/2808797.2809327.
20. **Huber, P., Oberdabernig, D. A. (2016).** The impact of welfare benefits on natives' and immigrants' attitudes toward immigration. European Journal of Political Economy, Vol. 44, pp. 53–78. DOI: 10.1016/j.ejpoleco.2016.05.003.
21. **Ignatow, G., Mihalcea, R. (2017).** An introduction to text mining: Research design, data collection, and analysis. SAGE Publications, Inc.
22. **Ignatow, G., Mihalcea, R. (2017).** Text mining: A guidebook for the social sciences. SAGE Publications, Inc. DOI: 10.4135/9781483399782.
23. **Jeong, Y., Suk, J., Hong, J., Kim, D., Kim, K. O., Hwang, H. (2018).** Text mining of online news and social data about chatbot service. Communications in Computer and Information Science, pp. 429–434. DOI: 10.1007/978-3-319-92270-6_61.
24. **Jia, S., Gao, L., Gao, Y., Nastos, J., Wen, X., Zhang, X., Wang, H. (2017).** Exploring

triad-rich substructures by graph-theoretic characterizations in complex networks. Physica A: Statistical Mechanics and its Applications, Vol. 468, pp. 53–69. DOI: 10.1016/j.physa.2016.10.021.

25. **Jonker, D., Brath, R. (2015).** Graph analysis and visualization: Discovering business opportunity in linked data. John Wiley and Sons.
26. **Kejriwal, M., Knoblock, C. A., Szekely, P. (2021).** Knowledge graphs: Fundamentals, techniques, and applications. The MIT Press.
27. **Kushwaha, A. K., Kar, A. K., Dwivedi, Y. K. (2021).** Applications of big data in emerging management disciplines: A literature review using text mining. International Journal of Information Management Data Insights, Vol. 1, No. 2, pp. 100017. DOI: 10.1016/j.jjimei.2021.100017.
28. **Liben-Nowell, D., Kleinberg, J. (2003).** The link prediction problem for social networks. Proceedings of the 12th International Conference on Information and Knowledge Management, pp. 556–559. DOI: 10.1145/956863.956972.
29. **Light, M. T., Thomas, J. T. (2021).** Undocumented immigration and terrorism: Is there a connection?. Social Science Research, Vol. 94, pp. 102512. DOI: 10.1016/j.ssresearch.2020.102512.
30. **Malak, M. S., East, R. (2016).** Spark GraphX in action. Manning Publications.
31. **Mihalcea, R., Radev, D. (2011).** Graph-based natural language processing and information retrieval. Cambridge University Press. DOI: 10.1017/cbo9780511976247.
32. **Miranker, M., Giordano, A. (2020).** Text mining and semantic triples: Spatial analyses of text in applied humanitarian forensic research. Digital Geography and Society, Vol. 1, pp. 100005. DOI: 10.1016/j.diggeo.2020.100005.
33. **Mukherjee, S., Oates, T., DiMascio, V., Jean, H., Ares, R., Widmark, D., Harder, J. (2020).** Immigration document classification and automated response generation. International Conference on Data Mining Workshops, pp. 782–789. DOI: 10.1109/icdmw51313.2020.00114.
34. **Negro, A. (2021).** Graph-powered machine learning. Manning Publications.
35. **Ott, R. L., Longnecker, M. T. (2015).** An introduction to statistical methods and data analysis. Cengage.
36. **Ozcan, S., Suloglu, M., Sakar, C. O., Chatufale, S. (2021).** Social media mining for ideation: Identification of sustainable solutions and opinions. Technovation, Vol. 107, pp. 102322. DOI: 10.1016/j.technovation.2021.102322.
37. **Raghavan, U. N., Albert, R., Kumara, S. (2007).** Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, Vol. 76, No. 3. DOI: 10.1103/physreve.76.036106.
38. **Robinson, I., Webber, J., Eifrem, E. (2013).** Graph databases. O'Reilly Media, Inc.
39. **Saura, J. R. (2021).** Using data sciences in digital marketing: Framework, methods, and performance metrics. Journal of Innovation and Knowledge, Vol. 6, No. 2, pp. 92–102. DOI: 10.1016/j.jik.2020.08.001.
40. **Sonawane, S., Kulkarni, P. A. (2014).** Graph based representation and analysis of text document: A survey of techniques. International Journal of Computer Applications, Vol. 96, No. 19, pp. 1–8. DOI: 10.5120/16899-6972.
41. **Witten, I. H., Frank, E., Hall, M. A. (2011).** Data mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc. DOI: 10.1016/c2009-0-19715-5.

Article received on 15/04/2023; accepted on 18/04/2024.
** Corresponding author is Esteban Castillo.*