Asesorías y Tutorías para la Investigación Científica en la Educación Puig-Salabarría S.C.
José María Pino Suárez 400-2 esq a Lerdo de Tejada. Toluca, Estado de México. 7223898475
RFC: ATI120618V12

Revista Dilemas Contemporáneos: Educación, Política y Valores.
http://www.dilemascontemporaneoseducacionpoliticayvalores.com/
Año: VII Número: Edición Especial Artículo no.:64 Período: Octubre, 2019.

TÍTULO: Proponer un nuevo método para la investigación de la relación entre las causas de divorcio y las reglas de asociación.

AUTORES:

1. Ph.D. Fatemeh Eghrari Solute.
2. Ph.D. Mohsen Yoosefi Nejad.
3. Ph.D. Mehdi Hosseinzadeh.
4. Ph.D. Nima Jafari Navimipour.
5. Ph.D. Amir Sahafi.
6. Ph.D. Aso Darwesh.

RESUMEN: Hoy en día, la minería de opinión juega un papel importante en el descubrimiento del conocimiento. Este tipo de conocimiento es un subconjunto de la minería de datos, y la minería de opinión se puede mejorar utilizando algoritmos de minería de datos. El análisis de sentimientos es una de las formas más importantes para la minería de opinión, donde millones de personas están expresando sus opiniones. Este estudio tiene como objetivo mostrar la relación entre las causas de divorcio utilizando minería de datos y reglas de asociación. Inicialmente, las principales causas de divorcio se examinaron desde la perspectiva de los usuarios, y luego, se indicó la relación entre estas causas para determinar las causas que conducen principalmente al divorcio. Los resultados de la simulación y la comparación en el conjunto de datos de divorcio revelan que el método propuesto tiene un rendimiento deseable.

PALABRAS CLAVES: minería de opinión, minería de datos, reglas de asociación, causas de divorcio.

TITLE: Proposing a New Method for Investigation of the Relationship between Divorce Causes and Association Rules

AUTHORS:

1. Ph.D. Fatemeh Eghrari Solute.
2. Ph.D. Mohsen Yoosefi Nejad.
3. Ph.D. Mehdi Hosseinzadeh.
4. Ph.D. Nima Jafari Navimipour.
5. Ph.D. Amir Sahafi.
6. Ph.D. Aso Darwesh.

ABSTRACT: Nowadays, opinion mining plays a major role in knowledge discovery. This kind of knowledge is a subset of data mining, and opinion mining can be improved using data mining algorithms. Sentiment analysis is one of the most important ways for opinion mining where millions of people are expressing their opinions. This study aims to show the relationship between divorce causes using data mining and association rules. Initially, the main divorce causes were examined from the perspective of the users, and then, the relationship between these causes was indicated to determine the causes which mainly lead to divorce. The results of simulation and comparison on the divorce dataset reveal that the proposed method has a desirable performance.

KEY WORDS: Opinion Mining, Data Mining, Association Rules, Divorce Causes

INTRODUCTION.

Opinion mining means natural and analytic-textual language processing in order to discover and extract the mental quantities of text sources. Generally, opinion mining is a kind of sentiment analysis to prove the polarity of the source text (for example, to distinguish between negative, neutral and positive beliefs). Moreover, it includes identifying the degree of objectivity and subjectivity of a text (i.e., identifying fact data in opposition to opinions), sometimes known as opinion extraction. Opinion mining also means discovering and summarizing the explicit opinions about the selected capabilities of the products. Some authors call this as the sentiment analysis. All these three definitions can benefit from a large extent of the duplicated data provided by social networks.

This study proposes a new method for investigation of the users' opinions on divorce, consisting of several stages that uses using tools such as hashtags, tags, likes and so on. In this way, users are divided into different categories, and divorce causes are determined based on the users' comments. This paper is organized as follows: opinion mining, association rules, literature review (research records), proposed method and analysis of the simulation results.

DEVELOPMENT.

Opinion Mining.

Opinion mining is a way to know the perspectives of the different people on a certain subject (Malik et al, 2018). These opinions may be expressed directly or comparatively. Direct opinion about a subject shows the ideas or beliefs, however, analogical opinion is a kind of comparison. Furthermore, emotional expression can be classified according to sentence level, positive and negative characteristics, and so on (Moers et al, 2018).

Emotions are the result of a person's relationship with the environment, including complex sets and coordinated components for responsiveness. These responses may include physical regulations, emotional or practical expressions. In an individual's interaction with his or her environment, one may encounter with a series of problems in life. People have different degree of emotions; these emotions may be positive or negative when personal experiences interact with the environment (Raisa Varghese, Jayasree, 2013; Khushboo, 2016).

Association rules.

Data analysis to extract hidden patterns and the relationship between them is one of the most important issues in data mining. Data mining uses a variety of tools for extraction in large datasets, including machine learning, neural network, decision tree, Bayesian network and association rules. Association algorithms and rules have high validity in data mining and they are used to detect the hidden relationships and data communications in a transparent and precise way (Allahyari et al, 2017).]; for example, association rules are used to analyze the causes of success or failure of an issue. This study uses association rules for analyzing and discovering the relationship between divorce causes relying on the users' comments in a seemingly insignificant form of vast data (Yuan, 2017; Doshi & Joshi, 2018).

Literature review (research records).

To date, a high number of algorithms have been proposed for association rules. Searching for repetitive items is the main and relatively time-consuming part of the most of the algorithms such as Apriori as one of the basic and well-known methods.

After discovering all the repetitive items in the dataset, the production occurs directly and quickly; therefore, the only difference between various proposed methods is the way that repetitive items are

discovered. The proposed algorithms have different approaches to optimize this process. Most of the methods focus on decreasing the number of times of reading the data on the disk.

In (Arpita Lodha1, 2016), a new method based on Apriori was presented. In this method, other smaller sets are not computed, but the set of case rules of the largest set are started to be constructed, and when acceptable solutions are obtained, the program stops. This method is of very high speed. In this method, first, the largest sets of user-selected items are built, and these sets are analyzed in terms of two criteria of coverage and confidence, and if acceptable, other sets are not analyzed, otherwise, other sets are built and this process continues to find a strong set of rules. The performance of this algorithm is summarized as follows:

- Preprocessing and deleting unwanted or unnecessary data.
- Calculating the threshold.
- Selecting item by user.
- Building large sets.
- Analyzing the built set.

A new method for extracting association rules is presented (Peeyush Kumar, 2016). In this method, the user first selects one or more items and separates the algorithm of all the transactions that contain this item, and in the next step, analyzes only transactions that have a high threshold frequency.

Each arbitrary subset of a repetitive pattern is a repetitive pattern, in itself. The graphic-data structure is used in the proposed method, where the vertices are items, and the edges represent the items that are selected together and the weights of the edges are equal to the number of choices that these two items are selected together. The problem of discovery of all subgraphs of a graph and of all repetitive pattern discovery algorithms are included in NP problems.

The complexity of the Apriori algorithm depends on the number of records and the number of items, however, in the proposed method, the complexity of the algorithm is limited to the number of items or the number of vertices of the graph. In the proposed method, the number of sets, which are the candidate patterns, are much less than those of the candidate patterns produced by algorithms and other methods, such as Apriori. Since by pruning the graph initially and the absence of edges of that item for producing the answer sets, though there are a high number of it, it leads to decrease in the number of the produced candidate patterns, thereby decreasing the scrolling number of the dataset and increasing the speed.

Researchers have proposed a method for finding association rules of intuitive data through the completion of association rule extraction, along with an Apriori-based retrospective peer-reviewed review (Suresh, 2015). This method is guaranteed to discover causality in a multidimensional and massive data set. Also, in this method, selecting a set of controlled variables is key for discovering qualitative dependency rules.

The validity of a set of controlled variables in real world applications guarantees the quality of the discovered dependency rules. The results showed that this method is faster than two effective limits based on the methods of discovery of dependency relationships and includes a combination of variables. This study aims to design an improved advisor system that can overcome the problems of the group refinement system (the cold start and the sparse rating matrix) and improve the accuracy of these systems. Here, the content information of the items and statistical information of users are used as a rich source of data. In the first part, user clusters are formed using the data in the rating matrix and the clustering algorithm and then, after determining the users for each cluster, existing items are averaged in order to show the similarity between items. Besides, data for content information of the items are used to calculate the similarity of the items.

Two values are obtained, in the first step of the process, the data in the user-pencil matrix is used to select the appropriate advisor system. First, the rating matrix is examined to determine whether the user is new or not. If the user is considered as a newcomer, s/he will enter the second advisor system. If the user is not new, she will enter the first. In this system for calculating the similarity, data in rating matrix and content information of the items are used; however, since the rating matrix is sparse, data need preprocessing.

The proposed method.

The relationship between the divorce causes from the users' perspectives was extracted as follows in the present study:

- Preprocessing comments: comments are saved as standard and unrelated information and comments are deleted.
- Tagging the positive and negative comments: comments are examined to see if they are positive or negative; positive one confirms the subject and negative one rejects it (Amir Hamzah, 2016).
- Hashtag tagging: The hashtag is the divorce cause and comments are examined to determine the hashtags.
- Extracting the comments with more than one factor (hashtag): Since the relationship between divorce causes is explored, comments that have several hashtags are very important (Bologna & Hayashi, 2018).
- Extracting all comments that have only one divorce cause.
- Investigating the relationship between agents according to association rules and analyzing results according to data mining and association rules.
- Predicting the divorces that have a cause and other causes that may most probably lead to divorce.

Hashtag posts and posts for each user group are separated, data for each one is shown in Figure 1.

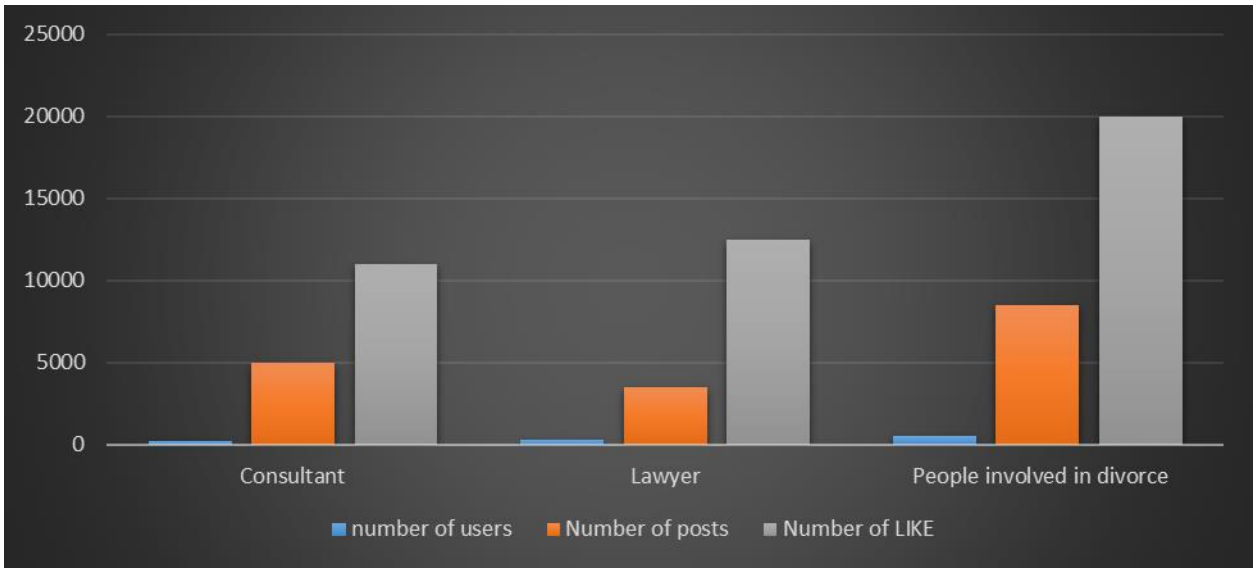

Figure 1. Content analysis based on user type.

Figure 2 shows the user's followers.

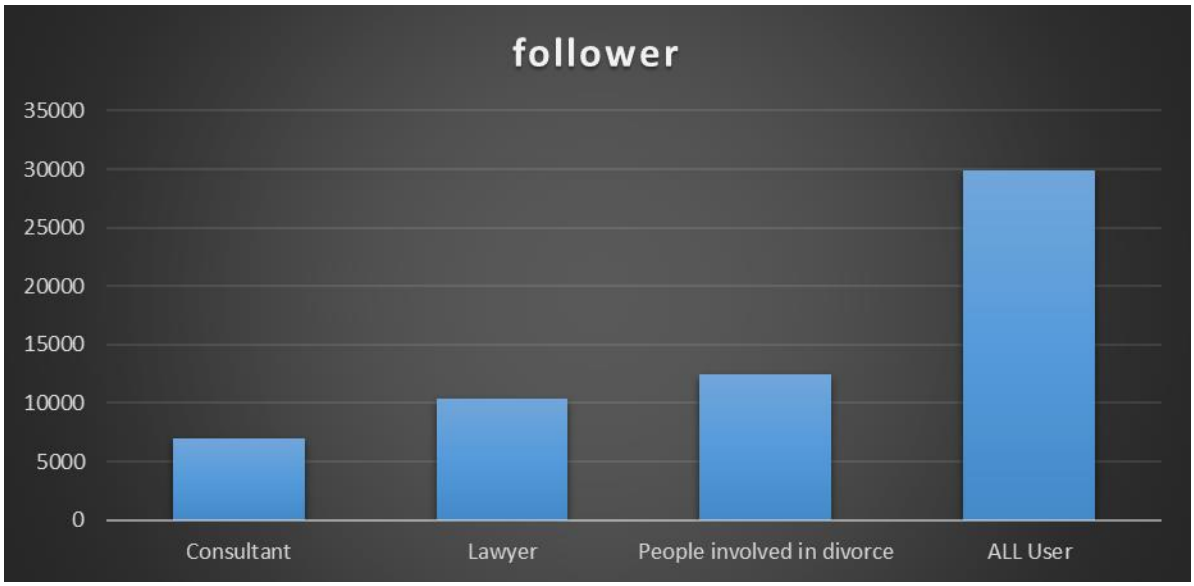

Figure 2. The number of filtered users' followers.

In Figure 3, content analysis of hashtags is visible to users (Fatemeh Eghrari Solout, Mehdi Hosseinzadeh, 2016).

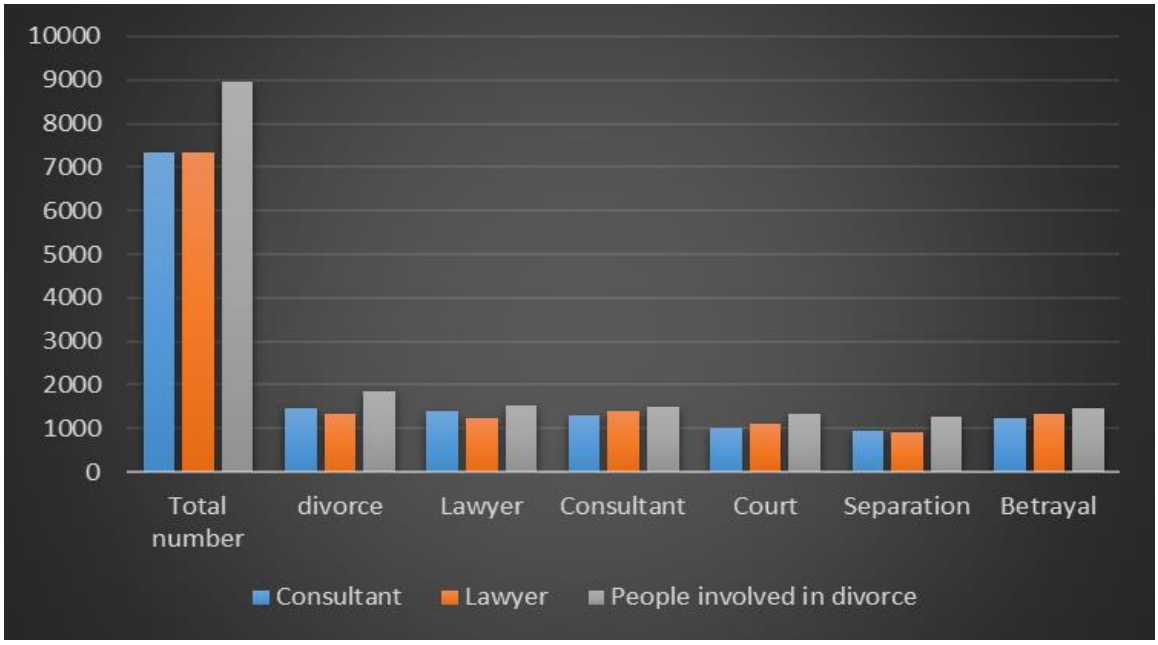

Figure 3. Hashtag content analysis based on users.

In this study, the causes of the divorce have been analyzed based on the users' comments and the tagging has been conducted in the same way. Tags in Table 1 have been used for tagging. These tags apply to the comments of the specified users.

Table 1. Divorce tags.

Divorce tag name.
Financial problem
Bad tempered
Unemployment
Sexual problem
Betrayal
Capricious
Addiction
Cold-tempered

Comments were categorized after tagging the comments based on the above tags, meaning each tag contains a set of comments. And any comment can express one cause for divorce. It should be checked that the comments are positive or negative, a positive statement is the one that has been confirmed by the user and confirms the sentence, but the negative sentence rejects it. The result of the tagged comment classification is visible in Table 2.

Table 2 the number of the comments of each tag.

Divorce tag name.	Number of comments.
Financial problem	3246
Bad tempered	2578
Unemployment	2045
Sexual problem	2162
Betrayal	3854
Capricious	2908
Addiction	2875
Cold-tempered	1366

In this section, comments have been re-tagged to distinguish positive and negative sentences. Tag set of the Intelligent Processing Institute has been used for tagging that contains 10 million dominant words on different tags, tagged with positive or negative.

After tagging, the comments were counted to show the percentage of the positive and negative sentences for each cause of the divorce. Initial data should be considered to count the likes of positive and negative comments. The result of this study shows the causes of the divorce from the user's perspective, which is visible in Table 6.

Table 3. Number of likes of positive and negative sentences of each tag.

Tag name	Number of comments.	Number of positive comments.	Number of positive like.	Number of negative comments.	Number of negative like.
Financial problem	3246	2041	9532	1753	7245
Bad tempered	2578	1057	7965	982	6421
Unemployment	2045	1164	8695	851	6308
Sexual problem	2162	1425	9337	635	7561
Betrayal	3854	2745	10267	843	8435
Capricious	2908	2078	9397	742	6718
Addiction	2875	1578	5463	953	6512
Cold-tempered	1366	523	6452	451	5178

In Figure 4, the results of the causes of divorce are visible.

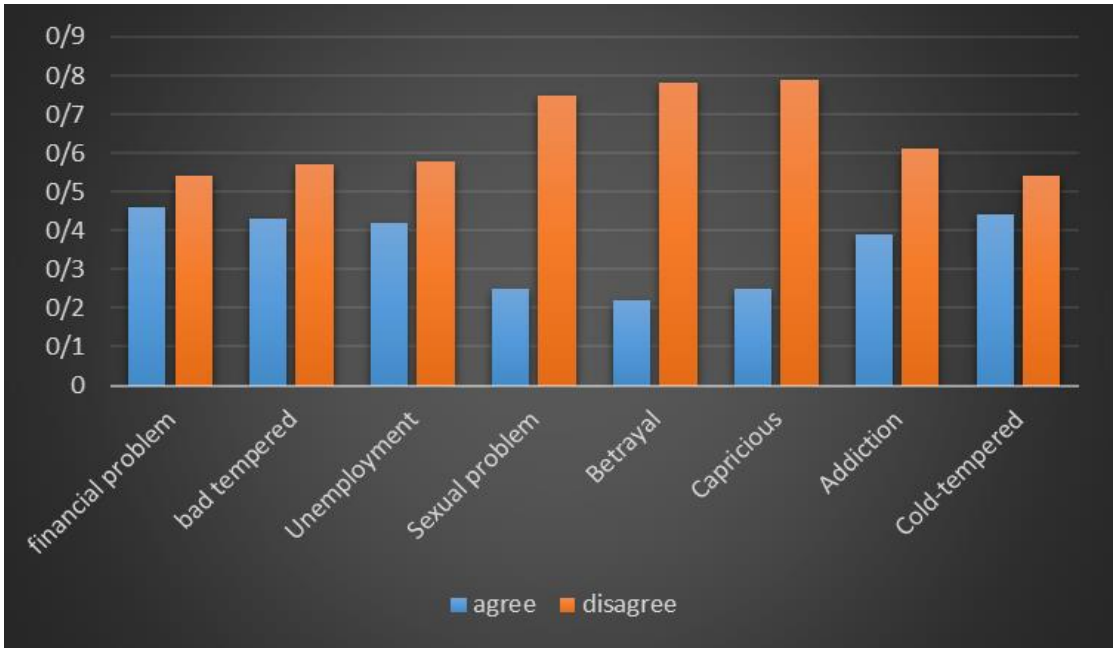

Figure 4 causes of divorce from the users' perspective.

For association rules, the relationship between the causes of divorce is analyzed as follows:

After the previous sections, the divorce dataset is built as shown in Table 4, of course, for separate positive and negative comments, in other words, for each data group, a certain data set is built as follows:

Table 4: Dataset format for extracting association rules.

	A	B	C	D	E	F
C1	1	1	0	1	1	1
C2	0	1	1	0	1	1
C3	1	0	1	1	0	1
C4	0	1	1	1	1	0
C5	1	0	0	1	0	*
C6	1	1	1	1	1	0

In Table 4, above c1, c2, c3.... shows a comment on divorce, and A, B, C ... each one is a hashtag, and the value of 1 indicates the use of hashtag and the zero value shows the lack of use of hashtag.

For analysis of the relationship between hashtags, each of which is a cause of divorce, a target hashtag is selected, for example, hashtag A, and its relationship with other hashtags is investigated, which its dataset is visible in Table 5.

Table 5: Dataset for selected hashtag.

	A	B	C	D	E	F
C1	1	1	0	1	1	1
C2	1	0	1	1	0	1
C3	1	0	0	1	0	1
C4	1	1	1	1	1	0

A threshold limit is defined to show the effectiveness of the hashtags. It is equal to half the number of comments that have the certain hashtag that is considered 2 in this example.

First, the frequency of features is calculated.

{A}=4,{B}=2, {C}=2, {D}=4,{E}=2, {F}=3

Due to the fact that the frequency of all hashtags is higher than the threshold, no hashtag is deleted and the L1 set is built.

L1={A,B,C,D,E,F}

The two-member set is built using L1 that is a combination of all parts of L1, and is calculated after identifying the frequency of each set in the dataset.

{A,B}=2 {A,C}=2 {A,D}=4 {A,E}=2 {A,F}=3 {B,C}=1 {B,D}=2 {B,E}=2 {B,F}=1 {C,D}=2 {C,E}=1 {C,F}=1 {D,E}=2 {D,F}=3 {E,F}=1

Those datasets with frequency higher than threshold are included in L2.

L2={{A,B},{A,C},{A,D},{A,E},{A,F},{B,D},{B,E},{C,D},{D,E},{D,F}}

Using the L2 set, the L3 set is built and their frequency is calculated.

{A,B,C}=1 {A,B,D}=2 {A,B,E}=2 {A,B,F}=1 {A,C,D}=2

{A,C,E}=1 {A,C,F}=1 {A,D,E}=2 {A,D,F}=3 {A,E,F}=1

{B,D,E}=2

{D,E,F}=1

L3={ {A,B,D} {A,B,E} {A,C,D} {A,D,E} {A,D,F} {B,D,E} }

The L4 set is built using the L3 set.

{A, B, D, E} = 2 {A, B, D, F} = 1 {A, D, E, F} = 1

L4 = {{A, B, D, E}}

The final set of L2, L3, L4 and L1 is as follows:

Answer={={{A,B}, {A,C}, {A,D},{A,E}, {A,F},{B,D}, {B,E}, {C,D},{D,E},{D,F}, {A,B,D} {A,B,E} {A,C,D} {A,D,E} {A,D,F} {B,D,E},{A,B,D,E} }

In the third phase, the confidence and coverage level for the above set should be calculated. In this calculation, the rule $A \rightarrow X$ must be checked for confidence and coverage and x is the elements of the answer set. In this example, the confidence level is 70%.

Following relationships are used to calculate the confidence and coverage.

- Support $(x \rightarrow y) = \frac{\text{support } (x \cup y)}{m}$
- $\text{Conf}(x \rightarrow y) = \frac{\text{support } (x \cup y)}{\text{Support } (x)}$

Analysis of results.

In this section, the described method has been analyzed and the results of the implementation of the proposed method and the rules have been described. Table 6 shows the extent of coverage and confidence for the extracted divorce causes; this Table shows the addiction as the main cause of the divorce, and the relationship between addiction and other causes has been demonstrated.

Table 6: Confidence and coverage of the rules.

Status	Coverage	confidence	Rule	Set
Weak	0.70	0.70	[addiction > cold-tempered]	[cold-tempered, addiction]
Weak	0.65	0.62	[addiction > capricious]	[capricious, addiction]
Strong	0.89	0.85	[addiction > betrayal]	[betrayal, addiction]
Weak	0.68	0.5	[addiction > sexual problem]	[unemployment, addiction]
Strong	0.79	0.83	[addiction > unemployment]	[unemployment, addiction]
Weak	0.70	0.71	[addiction > cold-tempered, unemployment]	[addiction, unemployment, cold-tempered]
Weak	0.33	0.5	[addiction > financial problem, bad-tempered]	[addiction, financial problem, bad-tempered]
Weak	0.72	0.65	[addiction > sexual problem, betrayal]	[sexual problem, betrayal, addiction]
Weak	0.70	0.70	[addiction > capricious, unemployment]	[capricious, addiction, unemployment]
Strong	0.66	0.75	[addiction > unemployment, betrayal]	[unemployment, betrayal, addiction]
Weak	0.65	0.67	[addiction >cold-tempered, financial problem]	[addiction, financial problem, cold-tempered]

For all the causes mentioned in this study, a table is plotted as above, and the relationship of each cause with other causes is discovered. In Figure 5, those causes that have a high level of coverage are visible.

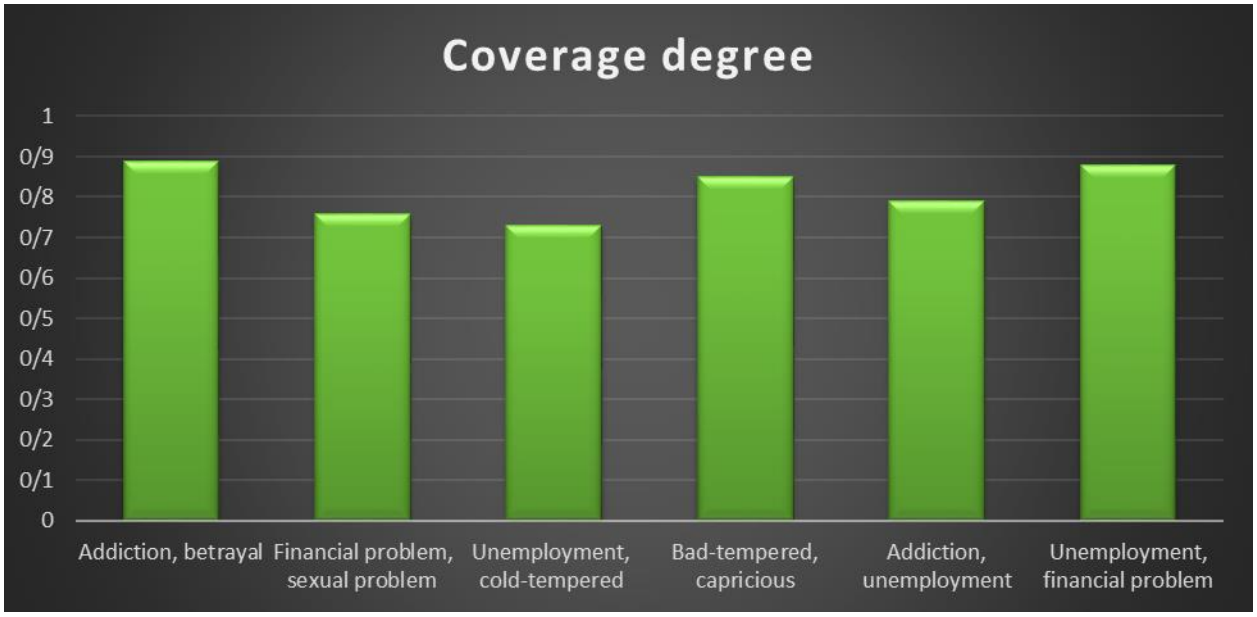

Figure 5. Coverage degree for different causes.

In Figure 6, those causes that have a high degree of confidence are shown.

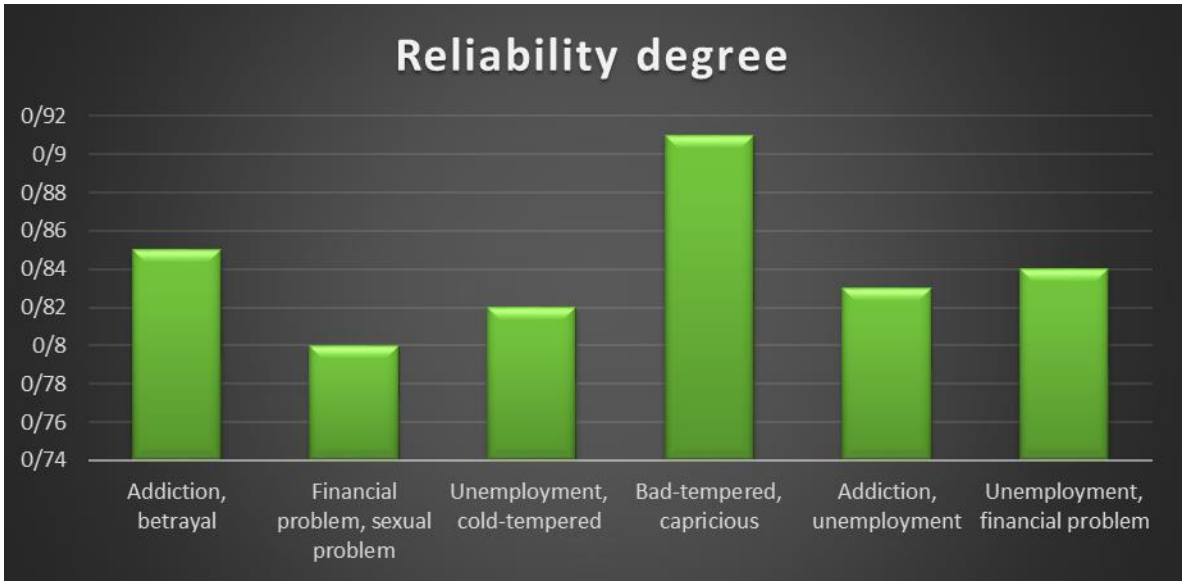

Figure 6. Confidence degree for different causes.

CONCLUSIONS.

This paper addresses the opinion mining on divorce and extracts the relationship between divorce causes using association rules.

Comments are analyzed based on a multi-stage approach based on hashtag, users, positive and negative sentences, and the number of likes. Based on these factors, divorce causes are analyzed from the perspective of the users. Subsequently, using association rules, it was determined that all causes together could affect the divorce. Two criteria of coverage and confidence were used to analyze the output of the paper.

BIBLIOGRAPHIC REFERENCES.

1. Malik, M., Habib, S., & Agarwal, P. (2018). A Novel Approach to Web-Based Review Analysis Using Opinion Mining. *Procedia Computer Science*, *132*, 1202-1209.
2. Moers, T., Krebs, F., & Spanakis, G. (2018, January). SEMTec: Social Emotion Mining Techniques for Analysis and Prediction of Facebook Post Reactions. In *International Conference on Agents and Artificial Intelligence* (pp. 361-382). Springer, Cham.
3. Raisa Varghese, Jayasree, (2013). "A Survey on Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Technology (IJRET), 2(11).
4. Khushboo, T. (2016) "Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm" ,Int. Journal. Computer Technology & Applications, page 211-216.
5. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.
6. Yuan, X. (2017, March). An improved Apriori algorithm for mining association rules. In *AIP Conference Proceedings* (Vol. 1820, No. 1, p. 080005). AIP Publishing.

7. Doshi, A. J., & Joshi, B. (2018). Comparative analysis of Apriori and Apriori with hashing algorithm.
8. Arpita Lodha1. (2016). A Modified Apriori Algorithm for Mining Frequent Pattern and Deriving Association Rules using Greedy and Vectorization Method, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization), 4(6).
9. Peeyush Kumar. (2016). A Review on Mining Frequent Patterns and Association Rules Using Apriori Algorithm, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization), 4(5).
10. P. Suresh. (2015). Improved Generation of Frequent Item Sets using "Apriori Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, 4(10).
11. Amir Hamzah. (2016). Opinion extracting and classification from questionnaire comments using HMM-POS Tagger and machine learning techniques, Data and Software Engineering (ICoDSE), 2016 International Conference.
12. Bologna, G., & Hayashi, Y. (2018). A Rule Extraction Study from SVM on Sentiment Analysis. *Big Data and Cognitive Computing*, *2*(1), 6.
13. Fatemeh Eghrari Solout, Mehdi Hosseinzadeh. (2016). Analysis of users' comments about the divorce factors, Journal of Advances in Computer Engineering and Technology, 2(4).

DATA OF THE AUTHORS.

1. **Fatemeh Eghrari** solute. Department of computer, Qeshm international Branch, Islamic Azad University, qeshm, Iran. Email: Eghrari.f@iau-qeshmint.ac.ir
2. **Mohsen Yoosefi Nejad.** Computer Engineering and Information Technology, Payame Noor University, Tehran, Iran. Email: m_yoosefi@pnu.ac.ir

3. **Mehdi Hosseinzadeh.** Health Management and Economics Research Center, Iran University of Medical Sciences, Tehran, Iran. Computer Science, University of Human Development, Sulaymaniyah, Iraq. Email: Hosseinzadeh.m@Iums.ac.ir
4. **Nima Jafari Navimipour.** Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran. Email: Jafari@iaut.ac.ir
5. **Amir Sahafi.** Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran. Email: a_sahafi@azad.ac.ir
6. **Aso Darwesh**. Information Technology Department, University of Human Development, Sulaimaniyah, Iraq. Email: aso.darwesh@uhd.edu.iq

RECIBIDO: 2 de septiembre del 2019. **APROBADO:** 10 de septiembre del 2019.