SpeD 2017
Bucharest, Romania
July 6-9, 2017

The 9th Conference on Speech Technology and Human-Computer Dialogue
July 6-9, 2017 – BUCHAREST, ROMANIA

Word Associations in Media Posts Related to Disasters - A Statistical Analysis

Mironela Pirnau
Faculty of Informatics - Titu Maiorescu University,
mironela.pirnau@utm.ro

Contents

1. Introduction
2. The analysis of the tweets frequency based on statistical indicators
3. Analysis of the groups of words
4. Descriptive statistics
5. Conclusions

1. Introduction

Aims:

- to analyze the frequency of the posts in case of earthquakes;
- to identify word associations Social Media posts;
- to identify the variation of a number of posts with unique content for the earthquake subject;
- to use the A-priori algorithm to extract words associations from these posts.

The present study uses messages generated by the Twitter platform, such as:

- *Vrancea (24. 09. 2016);*
- *Ussita (30. 10. 2016);*
- *New Zealand (13.11. 2016);*
- *Papua (23.01.2017).*

2. The analysis of the tweets frequency based on statistical indicators

- This paper is based on the following papers and studies for the analysis of the posts collected from Twitter in case of disaster or earthquake [bibliography];
- The tweets that were managed in order to carry out this study were collected by means of API streaming;
- This application, representing the basis of collecting the posts in case of disaster or earthquake, has been implemented in collaboration with Professor H.N. Teodorescu.

This application determines:

- For the study of the frequency analysis of the messages with the same content, only the data collected between 19th and 25th of September 2016 was used;
- If a tweet was encountered several times, the number of appearances was computed, so a new database with 1089 records, containing fields "unique_messages" and "frequency_ occurrence" was obtained.

Were identified:

- the vector comprising the tweets frequency – F (its elements must be ordered in an increasing manner so that each element observes the condition:
- $f_i < f_{i+1}$;
- the vector containing the number of times a tweet occurred - Vf.

Values found in the analyzed posts:

$$F \ = \{1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 17, 19\ , 20, 22,24,25\}$$

$$Vf \ = \ \{93, 870, 3, 8, 2, 3, 2, 2, 2, 1, 1, 4, 2, 1, 1, 2\}$$

For the experimental data in the vector F tweets frequencies, the following were computed:

- range of the data → Range=f_{max}-f_{min}
- Number of classes k using the H.A. Sturges' formula , $k = 1 + 3.322\log n$.
- The approximate class interval size → $h = \frac{Range}{k}$
- absolute frequencies;
- the relative frequency Fri (their sum is 1);
- increasing and decreasing cumulative frequencies.

3. Analysis of the groups of words

Were generated:

- The words association within the collected messages for a real earthquake occurrence;
- The frequency of words and their number corresponding to the tweet;
- For each words association found, the number of occurrences;
- The frequency of occurrences of a certain word association in the analyzed targeted tweets;
- The vector of the word association frequencies was ordered decreasingly.

A. Groups of words for the earthquake on 24th September 2016/ Vrancea

C1= {cutremur, seism, Vrancea, Buzau, Romania, victim, mort, mare, puternic, alerta, Richter, magnitudine}.

As a result, 46 word pairs were identified, and the word pairs having the occurrence frequencies >0.004 can be found in vector PW, and their frequencies in vector FW.

- **PW**={cutremur mare; cutremur Romania; cutremur seism; cutremur puternic; cutremur Vrancea; cutremur Richter; cutremur puternic seism; cutremur seism Vrancea; cutremur Vrancea magnitudine; cutremur magnitudine; cutremur puternic Romania; cutremur victim; cutremur Vrancea Richter};

- FW={0.043; 0.04; 0.037; 0.028; 0.024; 0.018; 0.017; 0.015; 0.011; 0.009; 0.008; 0.008; 0.007}

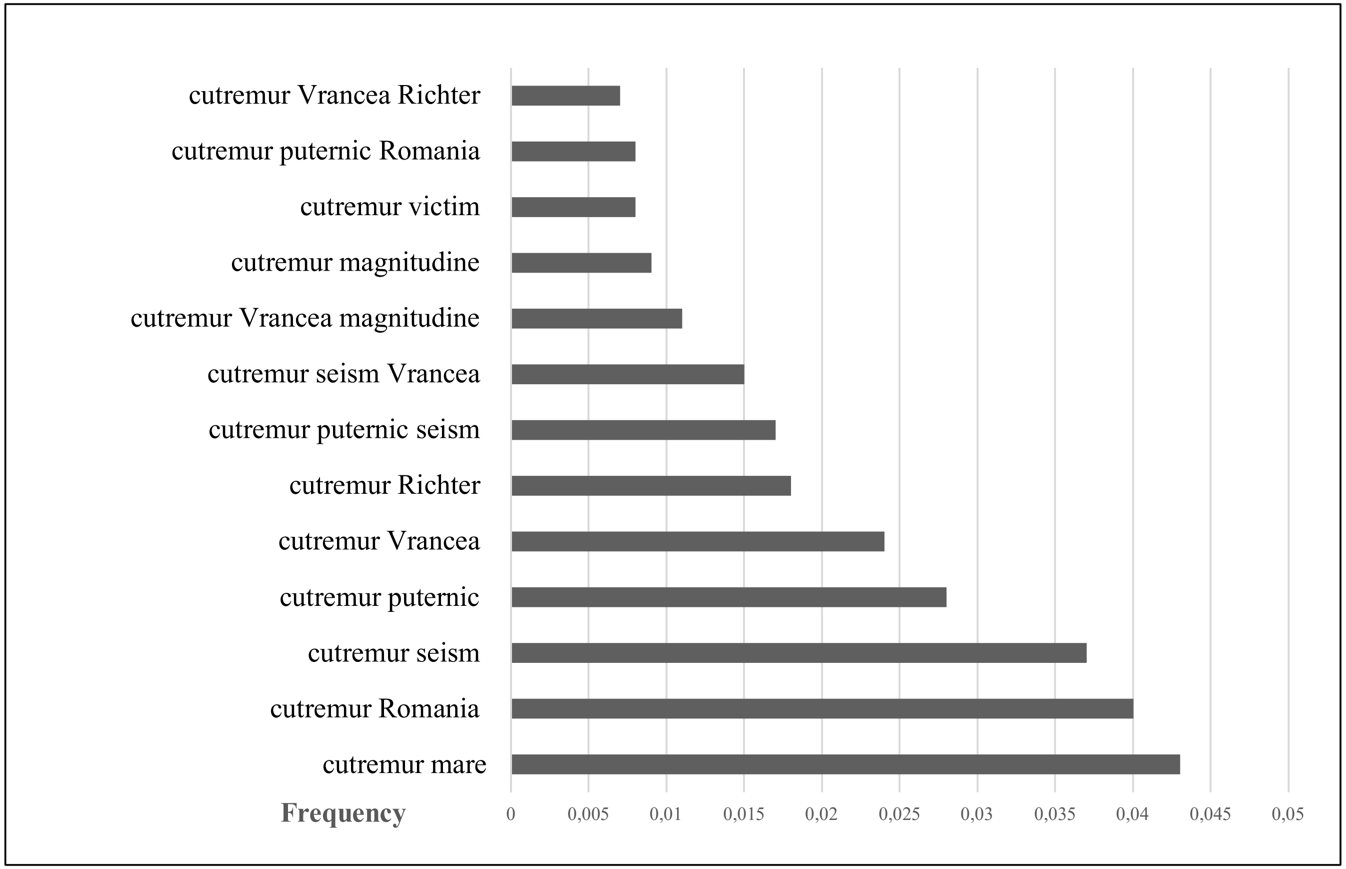


Fig. 1. Number of occurrences

The word **pairs** **”cutremur mare”, “cutremur Romania”, “cutremur seism”, “cutremur puternic” and “cutremur Vrancea”** are most commonly encountered.

The appearance frequency of the identified associations was computed using the elements presented in Fig. 1, relative to the total number of single tweets (1089 posts) containing the word "cutremur" .

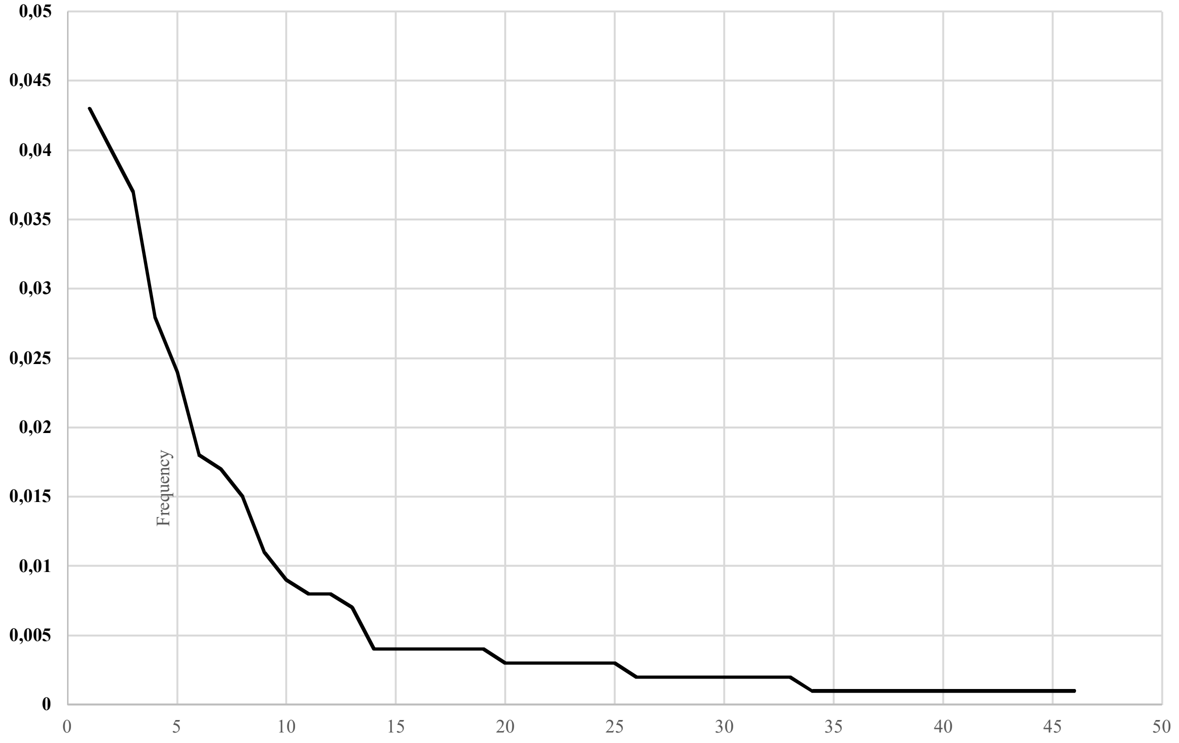


Fig. 2 Frequency of the identified words associations

B. Group of words for the earthquake occurred on 30th October/ Italy

Between 30th – 31 st October 2016 were extracted tweets, from which were kept for analysis a number of 39617, containing at least one derivation of strings from the elements in vector

”C2” = {**earthquake, terremoto, morta, Perugia, Ussita, vittima, attento, Richter, magnitud, Italy, dead, grande forte**}**.**

The frequencies of the association of the identified words

Frequency of word groups	*Word group associations/ tweets earthquake 30th October 2016, Ussita*
0.530	*terremoto Italy*
0.282	*earthquake Italy*
0.124	*magnitud earthquake*
0.119	*Italy earthquake magnitud*
0.027	*terremoto magnitud*
0.019	*terremoto Italy earthquake*

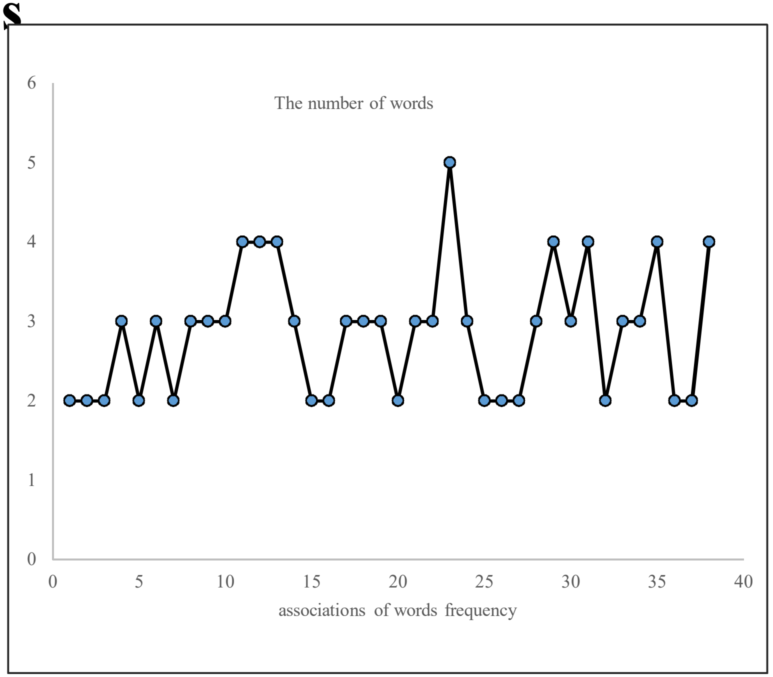


Fig. 3.The variation of the number of posts with a frequency <0.001

We found:

- **39 word associations with the appearance frequency between 0.001 and 0.53;**
- **54 associations with the appearance frequency between (0 – 0.001).**
- **The word pairs associations using the words with a very low appearance frequency from vector ”C2” – having values (0 – 0.001) Appendix.**

C. The earthquake on 13th November New Zealand 2016

A number of 19899 tweets which contain at least one word from the elements within vector ”C3” were kept for analysis

C3={**earthquake; tsunami; dead; Christchurch; Kaikoura; victim; Zealand;Richter; strong; seismo; magnitude; tragical; destroy**}.

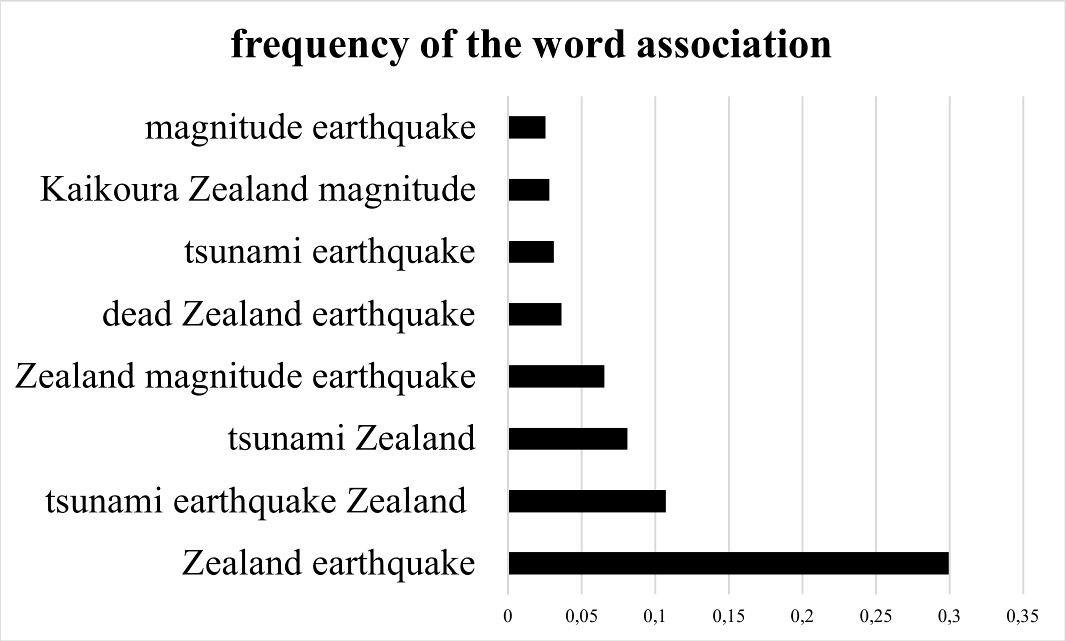


Fig. 4. Variation of the number of posts with frequency between [0.025 – 0.299]

The frequencies of the groups of five words from C3

Word Association	Frequency of word groups
tsunami Christchurch Zealand magnitude earthquake	0,003
tsunami dead Zealand magnitude earthquake	0,002
tsunami dead Christchurch Zealand earthquake	0,001
earthquake Christchurch Zealand strong magnitude	0,001
earthquake dead Zealand strong magnitude	0,001
earthquake tsunami Christchurch Kaikoura Zealand	0,0005
earthquake dead Christchurch Zealand magnitude	0,0004

There were identified six word associations corresponding to vector “C3” →{earthquake magnitude Tsunami dead Christchurch Zealand/ earthquake magnitude Tsunami dead Kaikoura Zealand/ earthquake magnitude tsunami strong Christchurch Zealand}.

D. Word associations for the earthquake on 22nd January 2017 in Papua

Partial - TABLE V. - WORD ASSOCIATIONS

Frequency of word groups	Word group associations/ tweets earthquake 22nd January 2017 Papua
0,3	magnitude earthquake
0,016	tsunami earthquake
0,012	Papua magnitude
0,011	Papua magnitude earthquake
0,007	victim earthquake
0,006	dead earthquake; tsunami Papua magnitude earthquake
0,004	strong earthquake

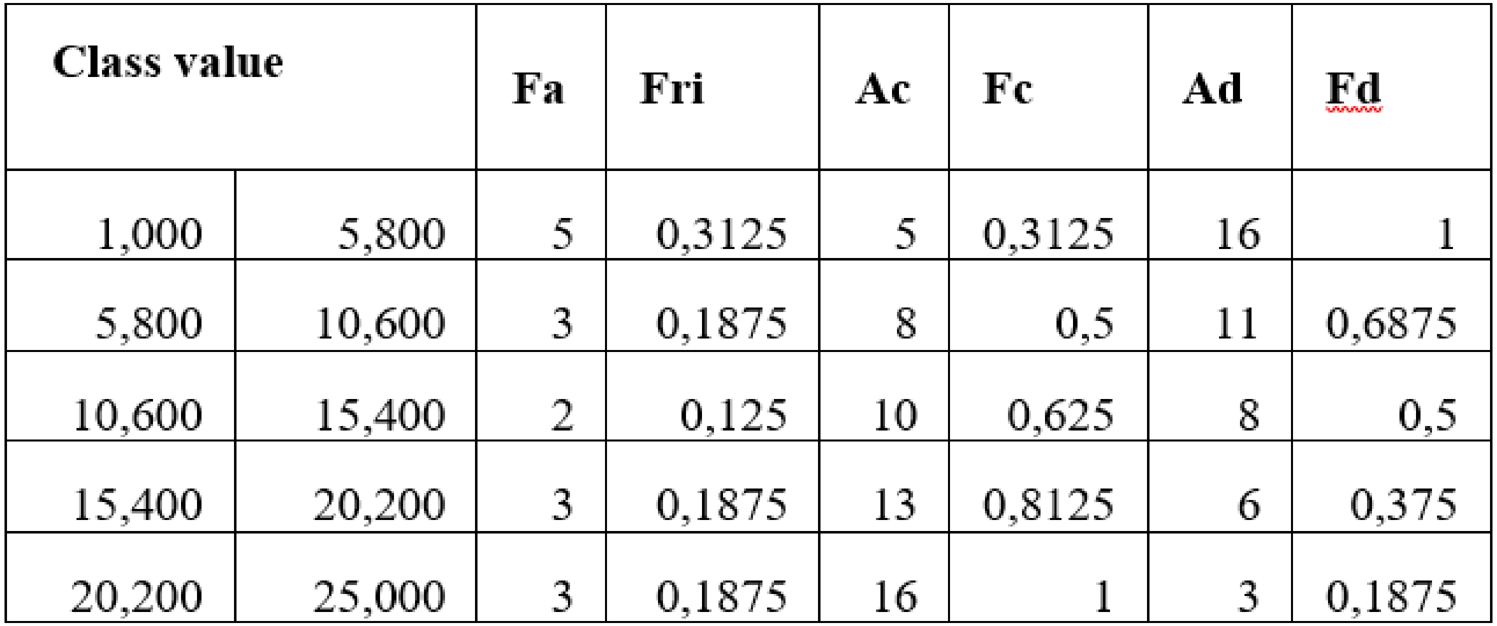
The analyzed tweets 16831 containing at least one element from vector "C4" were selected

= **{earthquake; tsunami; dead; Arawa; Bougainville; victim; Papua; Richter; strong; damaged; magnitude; tragically; destroyed}.**

4. Descriptive statistics

The number of classes and the absolute and relative frequencies for the number of tweets containing "cutremur" as the main word have been computed.

TABLE VI. CALCULATING THE RELATIVE AND ABSOLUTE FREQUENCIES



Class value		Fa	Fri	Ac	Fc	Ad	Fd
1,000	5,800	5	0,3125	5	0,3125	16	1
5,800	10,600	3	0,1875	8	0,5	11	0,6875
10,600	15,400	2	0,125	10	0,625	8	0,5
15,400	20,200	3	0,1875	13	0,8125	6	0,375
20,200	25,000	3	0,1875	16	1	3	0,1875

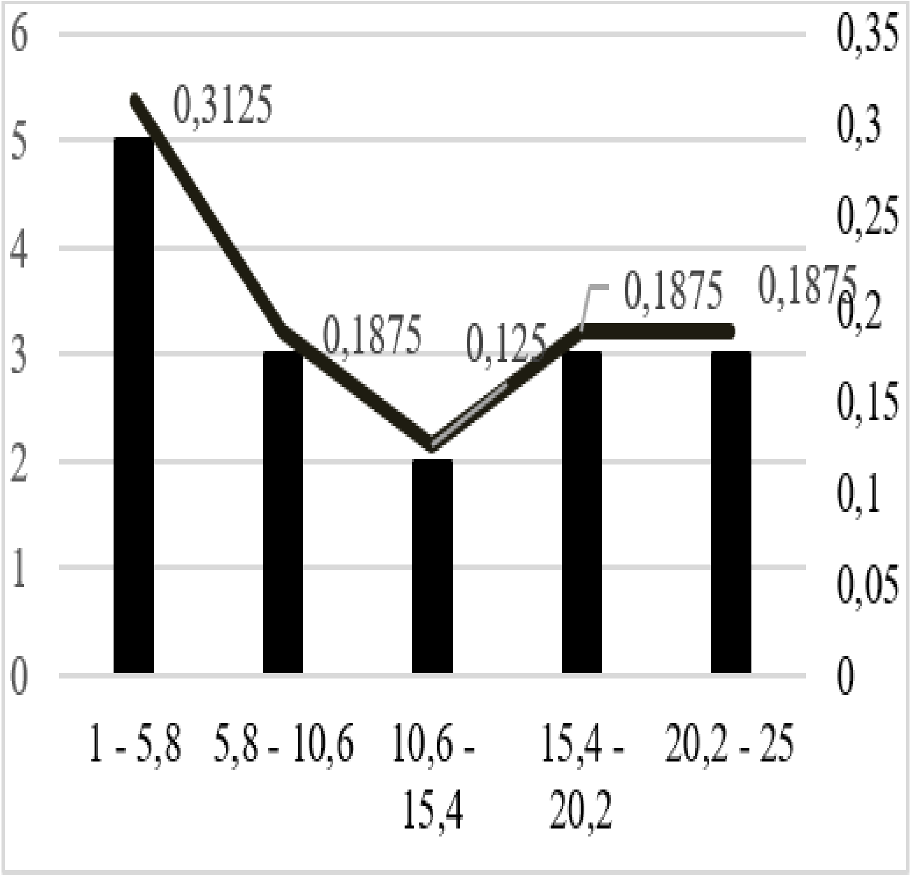


Fig. 6. Histogram of the absolute frequencies

TABLE VII. STATISTICAL INDICATORS

Statistical indicators	Number of words associated in one post		
	2	3	4
mean	22.9	5.388	2.058
median	23	3	2
mode	1	3	1
minimum	1	1	1
maximum	47	19	4
central value	24	10	2.5
amplitude	46	18	3
standard deviation	16.495	5.176	1.109
standard error	2.432	0.763	0.163
coefficient of variation	0.720	0.960	0.539
asymmetry	0.081	1.592	0.723
quartile 1	9.25	2	1
quartile 2	23	3	2
quartile 3	37.75	7	3

The data for these statistic determinations are given:

- $\text{Fm} = \{1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 17, 19, 20, 22{,}24{,}25\}$.
- the association frequency vectors of the analyzed words in section (iii) A, for two, three and four words were marked V[2], V[3] and V[4] and their values are:

V[2]= {47, 44, 40, 31, 26, 20, 10, 9, 1,1};

V[3]={19, 16, 12, 9, 8, 4, 4, 4, 3, 3, 3, 3, 2, 2, 2, 1, 1, 1 };

V[4]= {4, 4, 4, 3, 3, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1};

5. Discussion and Conclusions

- *The existence of the associations among the words in SM after a real earthquake occurrence .*
- *The importance of the discovery of new knowledge related to real events such as earthquakes.*
- *The criteria for analyzing this data depend on the collected data volume, as well as on the future effects of real events.*
- *By knowing the main word associations derivations of strings, that may occur within the SM posts in case of calamities, may lead to the timing optimization for analyzing their contents in real time.*

ACKNOWLEDGEMENT

I thank prof. H.N. (L.) Teodorescu for the idea of this research, for several personal communications and for his continuous support during the research.

Thank you for your attention!