4th International Conference on Automation, Control Engineering and Computer Science (ACECS - 2017)
Proceedings of Engineering and Technology – PET
Vol.19, pp.24-30

Get public opinion from the web: A survey of Text Mining techniques

Abdelkader RHOUATI *
Team SIQL, Laboratory LSEII, ENSAO,
Mohammed First University d'Oujda,
BP 669 Oujda. Morocco
abdelkader.rhouati@gmail.com

El Hassane ETTIFOURI
Team SIQL, Laboratory LSEII, ENSAO,
Mohammed First University d'Oujda,
Box 669 Oujda. Moroccoc
h.ettifouri@gmail.com

Mohammed G. BELKASMI
Team SIQL, Laboratory LSEII, ENSAO,
Mohammed First University d'Oujda,
BP 669 Oujda. Morocco
ghaouth@gmail.com

Jamal Berrich
Team SIQL, Laboratory LSEII, ENSAO,
Mohammed First University d'Oujda,
BP 669 Oujda. Morocco
jberrich@gmail.com

Toumi BOUCHENTOUF
Team SIQL, Laboratory LSEII, ENSAO,
Mohammed First University d'Oujda,
BP 669 Oujda. Morocco
tbouchentouf@gmail.com

ABSTRACT

The public opinion is a major factor in the political and economic decisions of all government. In fact, a respectful politician cannot ignore public opinion on making of any actions. However public opinion can be defined as a study of opinions, appreciations, and attitudes of individuals towards a particular subject. The arrival of web 2.0 over the last decade has changed the habits of the world. All people are now expresses its ideas and beliefs on the net through articles and comments. We are talking about public opinion on the Internet. Each website contains an enormous volume of texts representing public opinion. Getting this information requires automated extraction and analysis systems of those texts, we call this Text Mining. Over the past decade, a large amount research has been done to address this issue, hence the purpose of this work which represents a survey on the different Text Mining approaches and algorithms..

KEYWORDS

Public opinion, Text Mining, data analysis, NLP, dictionary, machine learning, semantic, statistical, POK platform

1 INTRODUCTION

Public Opinion refers to the opinion that is the most popular in community. It has changed form in last decade. People express their beliefs more and more on the internet by articles and comments on blogs and social networks. So to get public opinion from the internet, exactly from all texts written by people, we must analyze each text to determine the opinion orientation expressed: positive, negative or even neutral orientation. These processes are named Text mining. After that we can recover the global public opinion, which will be the sum of all discovered opinions orientations.

Basically, Text mining consists of three major steps:

- Text preprocessing: transform text into an intermediate form making easiest the analysis
- Text Analysis: extraction of knowledge from an analysis process.
- Results presentation

This work is carried out to make a survey of the main algorithms used by Text Mining process in the perspective to get public opinion from text published on the internet.

This article is organized in different sections. Section 2 details the general context of our work and give some important definitions, Section 3 and 4 introduce a categorization of the main algorithms used on Text Mining, Section 5 present a review of algorithms used in "Public Opinion Knowledge (POK)" platform. Finally, a conclusion and future works will be presented in Section6.

2 Context general: main definitions

2.1 Opinion definition

An opinion can be expressed about a person, an object, an action or even a topic of discussion. Basically, an opinion may concern everything. We speak of the target of opinion. We can formal the following definitions [1]:

Definition 1: The target of the opinion is called Entity (e), and it can be a material like a product, a book, a person, or even abstract as a service or a subject. So an entity is defined as a pair:

$$e = (T, W) \quad (1)$$

Where:

- T is a set of components and sub-components. It's important to notice that T can be also seen as an entity, and be defined as ***T = (T, W).***
- W is a set of attributes of e.

ISSN 2356-5608

In fact, this definition is a hierarchical one, which can be illustrated by the Fig. 1.

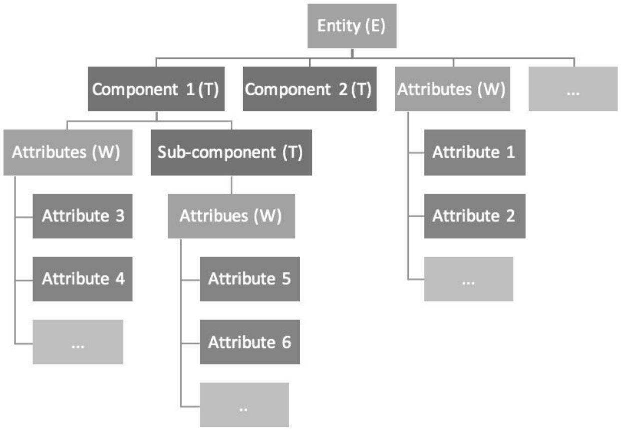


Figure 1: The hierarchical definition of entity

Definition 2: An opinion can be either general when it is expressed on the whole entity itself or regular when it concerns a part of the entity, such as attribute, a component or a sub-component. So an opinion is defined as a quintuple:

$$o = (e, a, oo, h, t) \tag{2}$$

Where:

- e is the name of entity,
- a is an aspect of e, which can be the whole entity, an attribute or even a component or subcomponent
- oo is the orientation of the opinion about a. this orientation can be positive, negative or neutral. And also can be expressed with different strength levels.
- h is the opinion holder
- t is the time when the opinion is expressed by h

2.2 Opinion public from the web

With the emergence of Web 2.0, the public opinion has been changed. We introduce this general definition:

Definition 3: The public opinion [2, 3, 4] refers to the opinion, positive or negative, that is the most popular. This opinion can be now an article on blogs, a comment on social networks, etc. Briefly any text content on the Web represents an opinion, and the whole consist the public opinion.

So the public opinion about an entity e can be formed as following.

Definition 4: in case of notation of opinion orientation with positive, negative or neutral, we can define the public opinion at time t by the following equation:

$$po_{et} = \sum_{i=1}^{nh} \sum_{j=1}^{na} \mathrm{oo}_{ij} \tag{3}$$

Where:

- “nh is the number of holder that gives an opinion of an aspect of the entity e.
- “na” is the number of aspect of the entity e.
- “oo_{ij}” the orientation of opinion. It takes three values (0 if opinion is neutral, 1 if positive and -1 if negative)
- “po_{et}” it the public opinion about entity e at time t. If po_{et} is upper than 0 then the positive opinion is the dominate, if it’s lower than 0 then it’s the negative opinion which dominates and if it equals to 0 then the public opinion is neutral about entity e.

In next chapter we will focus on a survey of approaches, methods and algorithms for analyzing a single text to retrieve an opinion orientation (oo).

3 Text Mining algorithms categorization

Text mining [5, 6] is a process of knowledge discovery from a text. In this Chapter we will focused on the core mining operations analysis that consist on pattern discovery, trend analysis and incremental knowledge discovery algorithms. So we will give a categorization of the main used algorithms. Various types of Categorization are possible, which depends on the main criterions of comparison between algorithms. Fig. 2 presents a summary of all algorithms categorization.

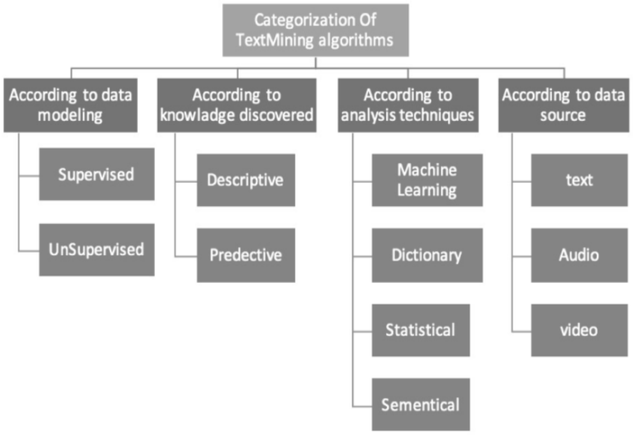

Categorization Of TextMining algorithms
According to data modeling
According to knowladge discovered
According to analysis techniques
According to data source
Supervised
UnSupervised
Descriptive
Predective
Machine Learning
Dictionary
Statistical
Sementical
text
Audio
video

Figure 2: Mapping of categorizations of data analysis algorithms

3.1 Supervised and unsupervised algorithms

This Categorization is based on how modeling the output and input data [7, 8].

1) *Unsupervised algorithms*

In "unsupervised" algorithms, the output conditions are not defined or represented in the dataset. The goal of such kind of algorithms is to uncover data patterns in the set of input fields.

2) *Supervised algorithms*

This categorization of algorithms use data in advance known class to which the data belong, and then constructs models. So they can use those constructed model to predict to which class and models the unknown data belong.

3.2 Categorization according to knowledge discovered

Text mining tasks can be classified into two categories: descriptive and predictive. This is can be done depending on the goal of exploration tasks [9].

1) Descriptive algorithms

A descriptive model, as his name said, allows to study important and various aspects of the data.

2) Predictive algorithms

The goal of a predictive model is to predict an unknown value, often in future, from exploration task of data.

3.3 Categorization according to technique used by mining analysis

This classification is according to used technique of the data analysis approach [10].

1) Machine Learning Approach

The Machine Learning Approach is an advanced technique that can be described as a two-step procedure;

Step1: learn the model from a corpus of data to be processed, either via supervised or unsupervised algorithms, and then enrich the basis of previously known models

Step2: classify the new data in the built model

2) Dictionary Approach

An algorithm based on a dictionary approach means that during the analysis of the text, words will be translated as a standard dictionary - word by word, usually without any correlation aspect between words.

3) Statistical Approach

Statistics is a part of text mining that provides analysis techniques to process large amounts of data which is interested in probabilistic models [11, 12].

4) Semantic Approach

Using a semantic [13] approach in text mining is very benefit in different ways, such as handling incomplete or erroneous text. The semantic algorithms study and observe the mechanisms proper to the construction of the meaning of the text, what we speak, what we want to say by words.

3.4 Categorization according to the type of data source

The data have now become enormous and diversified in a considerable way. A Categorization is then carried out according to the type of data processed in the data mining and text mining algorithms: Audio / video, text format.

4 Survey of the main Text Mining Algorithms

Text mining [14, 15] involves three major steps:

1/ Preprocessing of text to make storage of intermediate representation can make possible and simplify the next step.

2/ Analyzing these intermediate representations to extract knowledge

3/ Visualization of the results.

The categorization of Text Mining depends on the choice of algorithm and approach used of each steps. A part of our work consists to propose a map of the main used algorithms. so depending on the context of Text Mining we can used this map and choose the appropriate algorithms (In the extended version of this article we carry out a comparison of those main algorithms).

The Table. 1 represents our categorization of the main algorithms of Text Mining based on items of the Section 3.

5 POK platform: Review of used Algorithm

POK [41, 42], abbreviation of "Public Opinion Knowledge", it's a Big Data [43] platform to get the public opinion orientation about any subject from text content extracted from the web. The Fig. 3 illustrate the aim of POK Platform using Goal Question Metric (GQM) approach[44, 45].

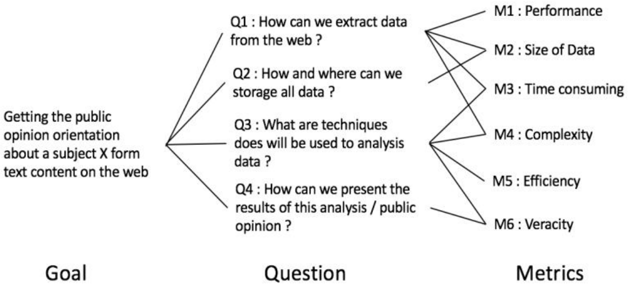
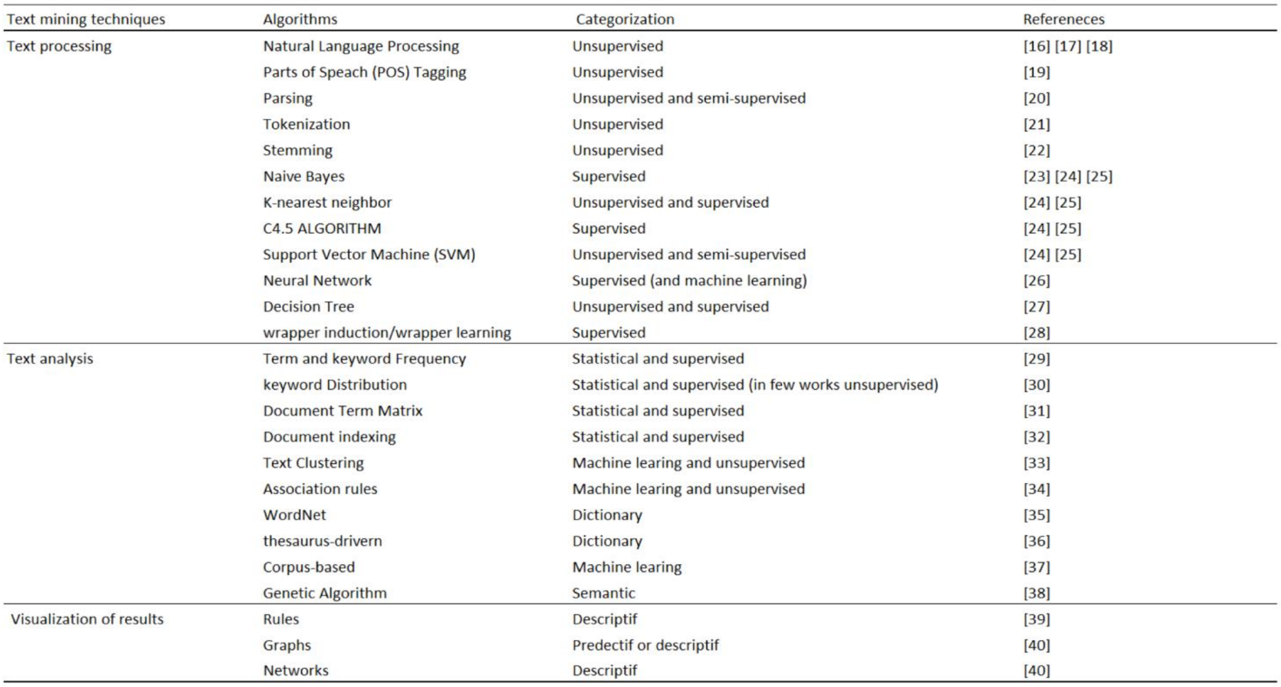

Getting the public opinion orientation about a subject X form text content on the web
Q1 : How can we extract data from the web ?
Q2 : How and where can we storage all data ?
Q3 : What are techniques does will be used to analysis data ?
Q4 : How can we present the results of this analysis / public opinion ?
M1 : Performance
M2 : Size of Data
M3 : Time consuming
M4 : Complexity
M5 : Efficiency
M6 : Veracity
Goal
Question
Metrics

Figure 3: Illustration using a GQM approach of the Problematic of discovering public opinion from text extracted from the web

Table 1: the proposed categorization of the main algorithms of Text Mining based on items of the Section 3



Text mining techniques	Algorithms	Categorization	Refereneces
Text processing	Natural Language Processing	Unsupervised	[16] [17] [18]
	Parts of Speach (POS) Tagging	Unsupervised	[19]
	Parsing	Unsupervised and semi-supervised	[20]
	Tokenization	Unsupervised	[21]
	Stemming	Unsupervised	[22]
	Naive Bayes	Supervised	[23] [24] [25]
	K-nearest neighbor	Unsupervised and supervised	[24] [25]
	C4.5 ALGORITHM	Supervised	[24] [25]
	Support Vector Machine (SVM)	Unsupervised and semi-supervised	[24] [25]
	Neural Network	Supervised (and machine learning)	[26]
	Decision Tree	Unsupervised and supervised	[27]
	wrapper induction/wrapper learning	Supervised	[28]
Text analysis	Term and keyword Frequency	Statistical and supervised	[29]
	keyword Distribution	Statistical and supervised (in few works unsupervised)	[30]
	Document Term Matrix	Statistical and supervised	[31]
	Document indexing	Statistical and supervised	[32]
	Text Clustering	Machine learing and unsupervised	[33]
	Association rules	Machine learing and unsupervised	[34]
	WordNet	Dictionary	[35]
	thesaurus-drivern	Dictionary	[36]
	Corpus-based	Machine learing	[37]
	Genetic Algorithm	Semantic	[38]
Visualization of results	Rules	Descriptif	[39]
	Graphs	Predectif or descriptif	[40]
	Networks	Descriptif	[40]

In this article we focus our discussion about Question 3; to be specific "what technique to be used to discover public opinion by analyzing a text?" it's clearly a matter of Text Mining, and the metrics 5 and 6 which are respectively efficiency and veracity. In fact, POK platform use a supervised algorithm to extract a known model data as article and comment from pages on the web, then it us a Natural Language Processing [46] to classify the text on an intermediate forms and finally it' use a dictionary approach to discover opinion orientation from the text. All those algorithms are implemented in a The Big Data platform Hadoop by implementing a distributed system based on a MapReduce technique. This POK's Text Mining approach is very difficult and has limits.

In one hand, a shallow of NLP techniques due to text word-level and Syntactic ambiguity of texts, the anaphora resolution and presupposition. Also the incapability to understand the context which is necessary to correctly interpret and understand any text [47]. On the other hand, whatever the relevance of the dictionary used in text mining algorithms, it will always be impossible to assemble all the words of the target language, especially due to various possible orthographic forms generated by the use of Tchat language or/and by the use of abbreviations to not exceed the limit of text characters on the web.

This will lead us to our future work. We plan to implement a new algorithm within the POK platform, which combines between a different technique: Semantic Natural Language for preprocessing [48] data, and machine learning dictionary [49] for text analysis. Then we will carry out a comparison study of the future results and the results of the current version of POK Platform.

6 CONCLUSIONS

In this paper we presented the general context and problematic of getting public opinion from text on the web with a focus on the core of text mining. We also provided detailed categorization and survey of the main exiting techniques and algorithms. So, in the second part we have discussed the text mining technique used in POK platform depending on categorization seen in the first part, as well as their limitations, which have lead us to our next work. It's consisting on the implementation of a new algorithm which combines different technique as semantic, machine learning and dictionary. And carry out a comparison study of the two results.

REFERENCES

[1] Bing Liu, Lei Zhang, "A survey of opinion mining and sentiment analysis" in Mining Text Data, pp 415-463, 2012. ISBN 978-1-4614-3223-4. Springer US.

ISSN 2356-5608

[2] Public opinion Quarterly journal, http://poq.oxfordjournals.org/, 2015.

[3] Mill, J. s. (2002 [1863]). On Liberty(see Chapter iV). in The Basic Writings of John Stuart Mill. new York: Modern Library; dewey, J. (1988 [1927]). The public and its problems. John Dewey: The later works, 1925–1953. Vol. 2, edited by J. Boydston. Carbondale: southern illinois university Press.

[4] Davison, W. P. (1958). the public opinion process. Public Opinion Quarterly, 22, 91–106; also see Allport, F. (1937). toward a science of public opinion. Public Opinion Quarterly, 1, 7–23.

[5] Neelamadhab Padhy, Dr. Pragnyaban Mishra, and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope" in International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[6] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan, "Data Mining, A Knowledge Discovery Approach", Springer, ISBN-13: 978-0-387-33333-5, 2007

[7] Cios, K.J., Pedrycz W., Swiniarski, R.W. & Kurgan, L.A. (2007), Data Mining: A Knowledge Discovery Approach, Springer, New York.

[8] Edin Osmanbegović, Mirza Suljić, "Data Mining approach for predicting student performance", Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.

[9] Nikita Jain, Vishal Srivastava, "Data mining techniques: a survey paper", IJRET: International Journal of Research in Engineering and Technology, pISSN: 2321-7308, Volume: 02 - Issue: 11, Nov-2013.

[10] Mikalai Tsytsarau, Themis Palpanas, "Survey on Mining Subjective Data on the Web", Survey on mining subjective data on the web. Data Mining and Knowledge Discovery 24: 478–514.

[11] JH Friedman, "Data Mining and Statistics: What's the connection?", Computing Science and Statistics, 1998 - venus.unive.it

[12] David J. Hand, "Data Mining: Statistics and More?", The American Statistician, volume 52, pages 112-118, 1998.

[13] Anna Stavrianou , Periklis Andritsos , Nicolas Nicoloyannis, Overview and semantic issues of text mining, ACM SIGMOD Record, v.36 n.3, September 2007, doi 10.1145/1324185.1324190.

[14] Ronen Feldman, James Sanger, "the text mining handbook", Cambridge university press, ISBN-13 - 978-0-511-33507-5, 2007.

[15] Aggarwal Charu C., Zhai ChengXiang, "Mining text data", Springer Science & Business Media, ISBN – 9781461432227, 2012.

[16] Fernando C.N. Pereira, Barbara J. Gross, "Natural Language Processing", ISBN 026266092X, Book 1994.

[17] Enrico Franconi, "Natural language processing", In The description logic handbook, Cambridge University Press, New York, NY, USA 450-461.

[18] Rus, Vasile, "Natural Language Processing", in Encyclopedia of Sciences and Religions, pages 1401-1404, isbn 978-1-4020-8265-8, 2013.

[19] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smit, "Part-of-speech tagging for Twitter: annotation, features, and experiments". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, USA, 42-47, 2011.

[20] Jay Earley, "An efficient context-free parsing algorithm". Commun. ACM 13, 2 (February 1970), 94-102.

[21] Benoît Habert, Gilles Adda, Martine Adda-Decker, Philippe Boula de Mareüil, Serge Ferrari, Olivier Ferret, Gabriel Illouz, Patrick Paroubek, "Towards Tokenization Evaluatio"n, in Proceedings First International Conference on Language Resources and Evaluation (LREC), Antonio Rubio, Navidad Gallardo, Rosa Castro, Antonio Tejada (resp.), Grenade vol. I, mai 1998, p. 427–431

[22] Julie Beth Lovins, "Development of a Stemming Algorithm" Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, March and June 1968

[23] Harry Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, FLAIRS Conference. AAAI Press, 2004

[24] Raj Kumar, Rajesh Verma, "Classification Algorithms for Data Mining: A Survey", in International Journal of Innovations in Engineering and Technology (IJIET), ISSN: 2319 – 1058, Vol. 1 Issue 2 August 2012.

[25] Xindong Wu et.al, "Top 10 Algorithms of Data Mining", Springer-Verlag London, 2007.

[26] Mark W. Craven, Jude W. Shavlik, "Using neural networks for data mining", Future Generation Computer Systems, Volume 13, Issue 2, 1997, Pages 211-229, ISSN 0167-739X.

[27] Chidanand Apté and Sholom Weiss, "Data mining with decision trees and decision rules". Future Gener. Comput. Syst. 13, 2-3 (November 1997), 197-210.

[28] Muslea, I., Minton, S. and Knoblock, C.A, "Wrapper induction by hierarchical data", Google Patents, Google Patents, 2003.

[29] Florian Beil, Martin Ester, and Xiaowei Xu, 'Frequent term-based text clustering', In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 436-442, 2002.

[30] Feldman Ronen, Dagan Ido, Hirsh Haym, "Mining Text Using Keyword Distributions", in Journal of Intelligent Information Systems, SN 1573-7675, 1998.

[31] Pauca, V. P., Shahnaz, F., Berry, M. W. and Plemmons, R. J., "Text Mining Using Non-Negative Matrix Factorizations". In SDM (Vol. 4, pp. 452-456), 2004.

[32] K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for Web document clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 10, pp. 1279-1296, Oct. 2004.

[33] Aggarwal Charu C and Zhai ChengXiang, "A Survey of Text Clustering Algorithms", in Mining Text Data, Springer US, isbn 978-1-4614-3223-4, 2012.

[34] Mahgoub H., Rösner D., Ismail, N. and Torkey FA, "text mining technique using association rules extraction". International journal of computational intelligence, 4(1), 21-28.

[35] George A. Miller, "WordNet: a lexical database for English", in Commun. ACM 38, 11 (November 1995), 39-41. DOI 10.1145/219717.219748, 1995

[36] Tsujii, Junichi, Ananiadou, Sophia, "Thesaurus or Logical Ontology, Which One Do We Need for Text Mining?", in Language Resources and Evaluation, N1 - V2, 1572-0218, 2005.

[37] Roger Garside, Geoffrey Leech, and Geoffrey Sampson, Review of "The computational analysis of English: a corpus-based approach", Comput. Linguist. 14, 4 (December 1988).

[38] J. Atkinson-Abutridy, C. Mellish, and S. Aitken, "Combining Information Extraction with Genetic Algorithms for Text Mining," IEEE Intelligent Systems, vol. 19, no. 3, 2004.

[39] Pak Chung Wong, P. Whitney and J. Thomas, "Visualizing association rules for text mining", in Information Visualization,. (Info Vis '99) Proceedings, IEEE Symposium on, San Francisco, CA, pp. 120-123, 152. doi 10.1109/INFVIS.1999.801866, 1999.

[40] D. A. Keim, "Information visualization and visual data mining," in IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1, pp. 1-8, Jan/Mar 2002. doi 10.1109/2945.981847, 2002.

[41] Rhouati Abdelkader, Ettifouri, El Hassane, Belkasmi, Mohammed Ghaouth, Bouchentouf, Toum, "Get the Public Opinion from Content Published on the Web/CSM: New Approach Based on Big Data" Proceedings of the Mediterranean Conference on Information Communication Technologies 2015: MedCT 2015 Volume 2

[42] Rhouati Abdelkader, Ettifouri, El Hassane, Belkasmi, Mohammed Ghaouth, Bouchentouf, Toum “Toward a Big Data Platform to Get Public Opinion from French Content on the Web/CMS", Europe and MENA Cooperation Advances in Information and Communication Technologies, International Publishing, isbn="978-3-319-46568-5", 2017.

[43] Min Chen , Shiwen Mao , Yunhao Liu, “Big Data: A Survey”, Mobile Networks and Applications, v.19 n.2, p.171-209, April 2014 , doi - 10.1007/s11036-013-0489-0.

[44] Basili, Victor; Gianluigi Caldiera; H. Dieter Rombach (1994). "The Goal Question Metric Approach"

[45] BASILI, V. 1994. GQM approach has evolved to include models. IEEE Softw. 11, 1, 8.

[46] Veronica Dahl, “Natural language processing and logic programming”, The Journal of Logic Programming, Volume 19, 1994, Pages 681-714, ISSN 0743-1066.

[47] C. D. Manning and H. Schutze, “Foundations of Natural Language Processing”, MIT Press, 1999

[48] Patrick Saint-Dizier, “An approach to natural-language semantics in logic programming”, The Journal of Logic Programming, Volume 3, Issue 4, 1986, Pages 329-356, ISSN 0743-1066.

[49] Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (March 2002), 1-47. DOI 505282.505283.

ISSN 2356-5608