The New Educational Review

Marius Cioca, Lucian-Ionel Cioca,
Cosmin Cioranu, Daniela Gifu
Romania

Extracting Features from the On-Line News for Making Templates Used in the Process of Educating the Next Generation of Politicians

Abstract

This paper presents an application for extracting features used in the process of education of politicians based on Open Source technologies, for extracting features from the public press; it enables the user to analyze data from text files and from the Internet. The analyzed data are from the field of politics, a topic which is both a current one and a cyclic process, occurring every four years for the parliamentary elections and every five years for the presidential elections. A specific political vocabulary was compiled for feature identification and analysis and for the interpretation of results obtained.

Keywords: *text classification, analysis, open source, political domain*

Introduction

Taking into account that the field of computers and information technology, as well as that of data volumes have grown exponentially, the use and development of new methods and technologies for revealing the information "hidden" in data have become imperative because such information cannot be detected by means of human analysis capacity (Hopcroft, J., Motwani, R., Ullman, J., 1979).

Communication is both a means and purpose to man, but to agents it is (still?) a means, and thus the starting point should be the object of communication, namely information (Barbat, 2002, p. 195).

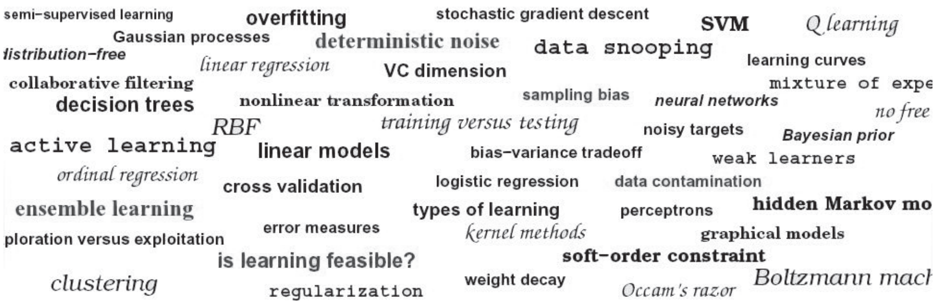
“Data” carries and contains this “information”, and thus, in order to reach it, we inevitably come to the so-called Data Mining (DM) (Filip, 2007), applied here to politics.

There are numerous definitions of DM, such as, for instance, the one provided by WordNet Search – 3.1, “*data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases; a way to discover new meaning in data*”.

This paper presents an application, developed in PHP (**H**ypertext **P**reprocessor) for extracting features from the public press; it enables the user to analyze data from text files and from the Internet (e.g. blogs, social networks, twitter, etc.). The analyzed data are from the field of politics (data from any field may be analyzed if a vocabulary specific to the analyzed field is provided), a topic which is both a current one and a cyclic process, occurring every four years for the parliamentary elections and every five years for the presidential elections (except for by-elections). A specific political vocabulary was developed for feature identification and analysis and for the interpretation of the results obtained.

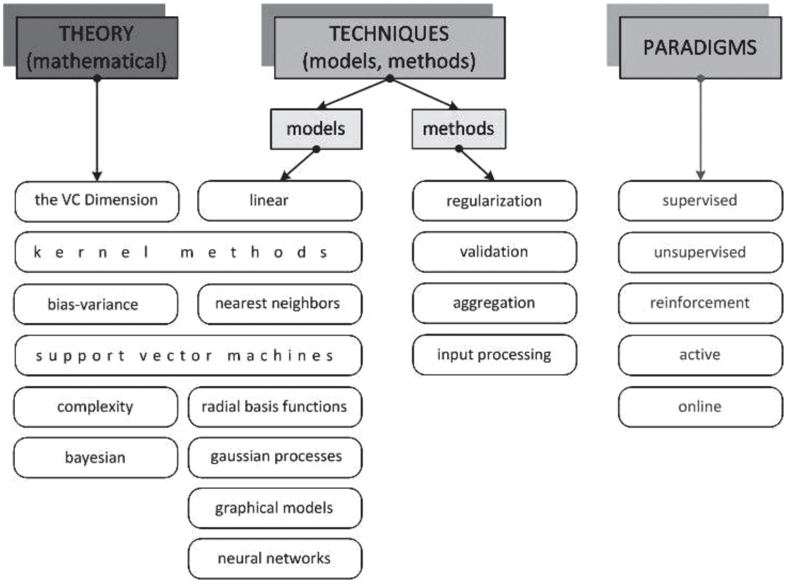
The Romanian and international DM and text analysis literature, published in print and on the Internet, is impressive. The multitude of solutions, theories, techniques (models, methods), paradigms, etc. is overwhelming. After a simple search on GOOGLE, we faced the situation shown in Figure 1 (Abu-Mostafa, Learning From Data, 2012).

Figure 1. It’s a jungle out there **(Abu-Mostafa, 2012)**



After a careful approach to the field and the accomplishments of Romanian and foreign scientists things became clearer and we came to the conclusion in Figure 2 adapted after Abu-Mostafa (Learning From Data, 2012).

Figure 2. Theories, techniques (models, methods) and paradigms used in DM – adapted after Abu-Mostafa (2012)



We analyzed the political discourse (in the public media) during the campaign in 2009 in order to develop profiles of the political leaders and to determine the preferences of the electorate. The main tool employed was the American software application LIWC2007 (LIWC: Linguistic Inquiry and Word Count, 2007), adapted for the Romanian language (whose parameter is the calculation of frequency – quantitative analysis), the results being afterwards interpreted qualitatively (qualitative analysis) based on graphical representations developed in Excel. At the beginning of the analysis, the main issue was to determine the categories (i.e. 28 classes, using their lexical families presented in Gifu, D. (2010) (Romanian lexical LIWC 2007) optimal for determining a certain *political attitude* able to influence the electorate's decision and which constituted the elements of the chart.

The results obtained in the process were satisfactory and are described in Gifu, D. (2010), Gifu, D. & Cristea, D. (2011) and Gifu, D. & Cristea, D. (2012); however, we identified two main drawbacks of LIWC 2007: a) We were able to analyze only text files, on-line data being inaccessible; b) There was no integrated graphic tool, and thus we had to use Excel.

Therefore, this study had three objectives:

- The development of a software application similar to LIWC2007, and dealing with the two drawbacks mentioned above (in a DM process we performed pre-processing of data);
- The use of our application for the analysis of the 2012 electoral campaign.

Methodology of Research

The vocabulary used by the application (28 classes), is almost identical to the one used in 2009, with slight "adjustments":

- All classes use only the root of the word taken from Dictionar explicativ al limbii romane | DEX online;
- The synonyms of the words that represent a certain class were taken from Dictionar de sinonime online.

Table 1. Example of 28 classes

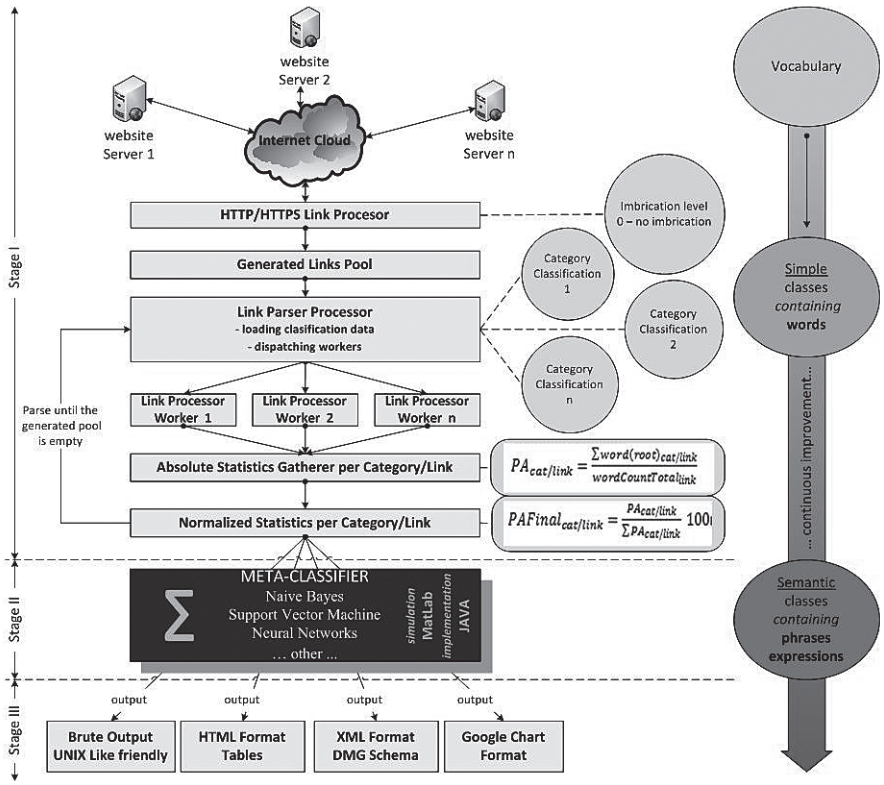
…	…	…	…
Social	…	Safety	Work
Family	…	…	…
…	…	…	…
…	Rational	…	…
Emotional	…	…	Financial
…	…	…	…

Table 2. Example of classes (10 words each)

Work	**Social**	**Financial (money)**
Qualification	Communication	Business
Organization	Friendly	Accounts
Advance	…	Estate
…	Consultation	Budget
Project	…	…
Work	Intervention	Auction
Activity	Event	Raw

Because all the classes (categories) defined and used by the application are balanced from a quantitative point of view (they contain between seven and ten words), in the application only the PAF normalization is performed (cf. Figure 3), namely a qualitative normalization. The LIWC2007 software application made a quantitative normalization, followed by a qualitative one (in Excel) because it considered the classes which compose the vocabulary as quantitatively disproportional. Moreover, the application uses a special class populated with "link words", which are not counted, in order to reduce the "noise" of results obtained during the data pre-processing stage.

Figure 3. The Application of Outline Scheme


website Server 2
website Server 1
Internet Cloud
website Server n
Vocabulary
HTTP/HTTPS Link Procesor
Imbrication level 0 – no imbrication
Generated Links Pool
Category Classification 1
Link Parser Processor - loading clasification data - dispatching workers
Category Classification 2
Category Classification n
Simple classes containing words
Parse until the generated pool is empty
Link Processor Worker 1
Link Processor Worker 2
Link Processor Worker n
Absolute Statistics Gatherer per Category/Link
$PA_{cat/link} = \frac{\sum word(root)_{cat/link}}{wordCountTotal_{link}}$
Normalized Statistics per Category/Link
$PAFinal_{cat/link} = \frac{PA_{cat/link}}{\sum PA_{cat/link}} 100$
... continuous improvement...
Stage I
Stage II
META-CLASSIFIER
Naive Bayes
Support Vector Machine
Neural Networks
... other ...
simulation
MatLab
implementation
JAVA
Semantic classes containing phrases expressions
Stage III
output
output
output
output
Brute Output UNIX Like friendly
HTML Format Tables
XML Format DMG Schema
Google Chart Format

The algorithm shown in Figure 3 is described below.

Requirements

a. The structure of primary elements, termed as word classes, specified in a "source word" format (the word without a prefix or suffix [the word root])

b. The link pool to be classified

Note: These sources are in standard configuration files (*.ini)

The Application Methodology

a. All categories are loaded in an accepted format
   - wordCountcat=[total no. of words / category]

b. Every link is downloaded
   - The algorithm does not generate links from a source document, i.e. if it finds a link it does not follow it; thus, the algorithm operates on the level of 0;

c. Tags are extracted

d. A structure based on the word is thus created, and all categories connected with that word are linked to it
   - It can also be checked whether there is at least one word in two categories and that word must be deleted from both categories, to prevent conflicts;

e. Each word is searched in the word source extracted and a distribution network is created in the following format
   - $[\mathrm{Word(root)}]_{\mathrm{cat/link}}$ = [number of occurrences in the linked document]

f. The absolute probability between the total number of words and the number of occurrences in a certain category is calculated

$$PA_{cat/link} = \frac{\sum word(root)_{cat/link}}{wordCountTotal_{link}}$$

g. The normalized probability is calculated

$$PAFinal_{cat/link} = \frac{PA_{cat/link}}{\sum PA_{cat/link}} 100$$

Output

a. Unix like configuration file – a format compatible with the large majority of programming languages, parsers are included at the API level, and thus it is a format which allows further cross-platform processing. The problem is that it does not allow for specifications of the types of fields used, and thus a wide range of interpretations remains at the level of source code of the data stored in this output format;

b. HTML Table – a format which ensures user interface, UI cross-platform compatibility, but it is difficult to work with it at an interoperability level of applications which use these output data;

c. XML – using DMG (DMG, 2012) ensures the best interoperability between applications which use this type of data. The advantage is due to the XML, which incorporates specifications about the types of data used in the output document, but also to the DMG specifications which provide a full understanding of the output data;

d. *Google API Charts* (Google.com) ensure visual comparison. Google Charts are integrated in the application using Javascript and HTML, which gives it versatility and enables it to clearly display results.

Results

The results obtained by our application (Figures 5 and 6), in comparison to LIWC2007 (Figure 4), are approximately identical. The data analyzed were those "sampled" during the 2009 campaign. Therefore, the first objective has been accomplished: the charts can be generated without using other applications (e.g. Excel) with the same results.

Moreover, Figure 7 shows the analysis of the data on two different Internet sites (this could be applied to any addresses), namely weblogs belonging to two Romanian political leaders, the President of the National Liberal Party (NLP), currently the ad-interim President of Romania (Crin Antonescu | Blogul Presedintelui PNL) and the personal blog of the leader of the Social Democrat Party, currently the Prime Minister of Romania (Blogul lui Ponta). This online analysis illustrates the second objective of this paper, which enables us to pre-process data directly on the Internet.

The interpreting of the data, among other aspects, answers a question which has hunted Romanians for a long time, "…how was it possible to create the Social Liberal Union (SLU), the political force currently governing Romania, by uniting two parties animated by totally opposed doctrines (i.e. the main parties which form the SLU: the SDP and the NLP)?"

Figure 4. Results obtained in 2009; the pre-processing was performed with LIWC2007, and the chart was made in Excel – from (Gifu, D., 2010)

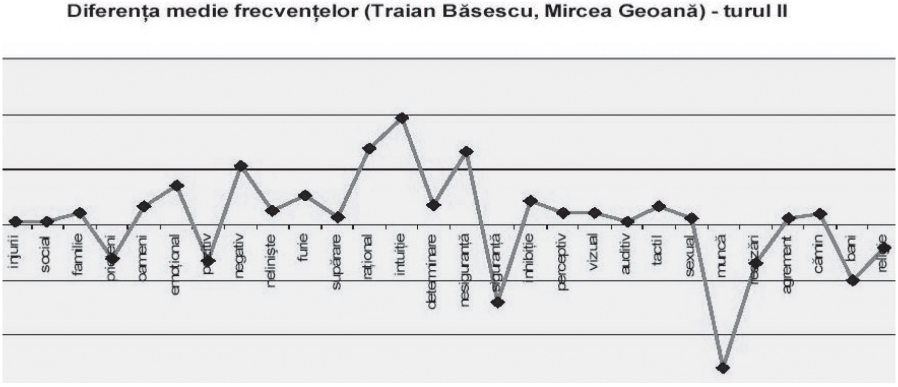


Figure 5. The results obtained on the same 2009 data with our own application; (line) graph generated with Google Chart integrated in the application with JavaScript and HTML

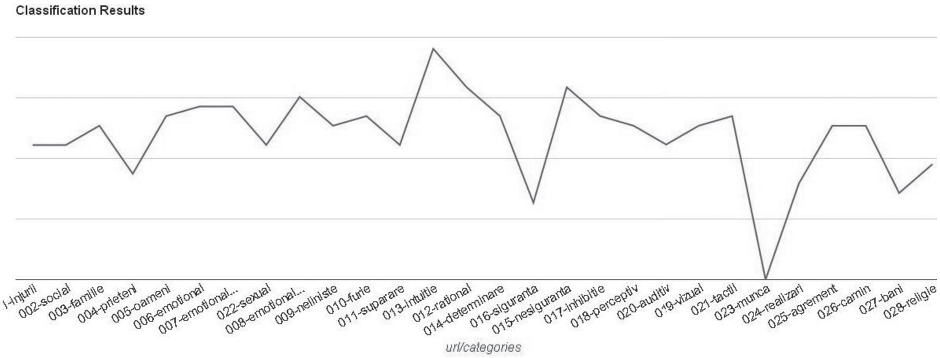
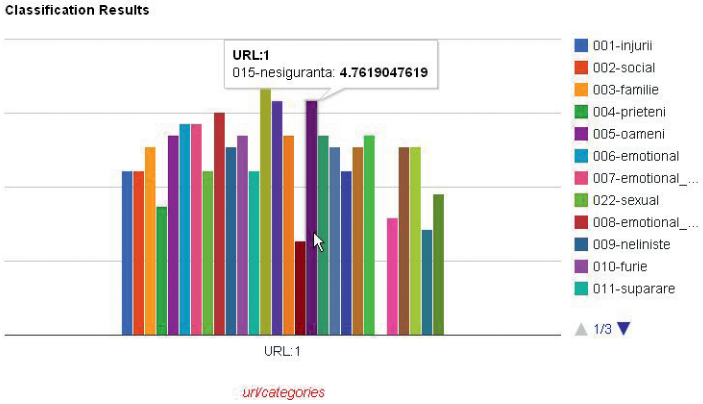
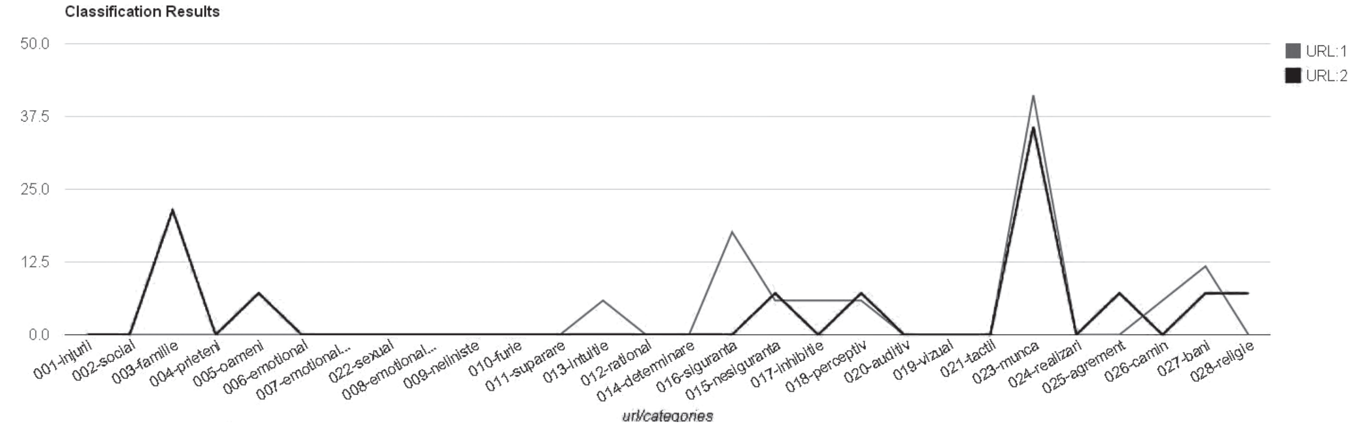


Figure 6. Results obtained on the same 2009 data with our own application; (tower) graph generated with Google Chart integrated in the application with JavaScript and HTML



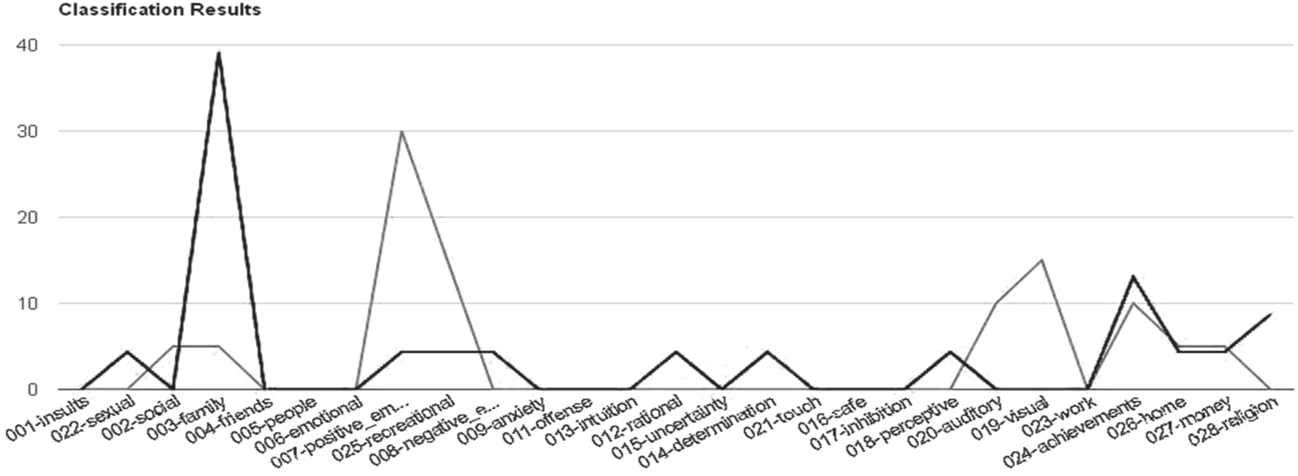
After analyzing the data (Figure 7), we managed to determine the mutual feature which laid the foundation for SLU, namely "work" ["munca" in Romanian] (as shown in the figure, this aspect is almost identical in the case of the two leaders). Other features specific to the both political leaders are the lack of aspects regarding: "strong language" ["injurii" in Romanian], "unrest" ["neliniste" in Romanian], "anger" ["furie" in Romanian]. The chart obtained also shows features specific of each of the two parties (even if they united, the two parties have preserved their

Figure 7. Graphic results obtained through the on-line analysis of the blogs belonging to the two political leaders (blue – the NLP leader and red – the SDP leader)



URL: 1: http://www.crinantonescu.ro/Blog/CrinAntonescu.html
URL: 2: http://blogponta.wordpress.com/

Figure 8. Graphic results obtained through the on-line analysis of the blogs belonging to the two political leaders (blue – Obama and red – Romney)



URL: 1: http://www.guardian.co.uk./word/2012/nov/07/barack-obama-speech-full-text

URL: 2: http://bostonherald.com/news/us_politics/view/202121107text__mittromneys_concession_speech identity), namely, the SDP has a social doctrine and is characterized by aspects such as, "family" ["familie" in Romanian], "people" ["oameni" in Romanian], "leisure" ["agrement" in Romanian] while the NLP is characterized by features such as "intuitive" ["intuitie" in Romanian], "safety" ["siguranta" in Romanian], "money" ["bani" in Romanian].

The results are presented after the analysis of the text in the Romanian language. The application analyses texts from any language if the dictionary is translated into that language. For exemplification, the same dictionary was translated into the English language and was analyzed on the candidates for US presidency (in their speeches for the 2012 elections). The results are presented in Figure 8, where we can notice and analyze the special features of both candidates, but at the same time, knowing the results of the elections, we can say the results presented in Figure 8 mirror the aspect of American people's requirements and expectations from their leaders, and the special features of this nation. The dictionary is carefully prepared, abundant with words and specific to particular domains, so the accuracy of the results increases.

Conclusions and Further Research

As shown in the outline scheme (Figure 3), the pre-processing of data was performed (stage I) as well as their representation in XML taking into account the DMG recommendations, and in a graphical form (tower, line) integrating into the application the possibilities provided by Google Chart (stage III).

Considering the fact that in a "knowledge-based society" we must develop "more intelligent" computers able to provide more consistent data, without daring to mention, yet, the concept of the "society of consciousness" launched by the late scientist and academician Mihai Draganescu in 2000 (Draganescu, 2000), in the future we aim to expand the "black box" in Figure 3 (stage II), turning towards Web 3.0, on the one hand through defining some semantic classes, and on the other through developing a meta-classifier providing solutions such as Naïve Bayes, Support Vector Machines, Neural Networks, etc. For simulating these solutions we shall use the development environment MatLab which quickly facilitates analyses such as Naïve Bayes (MathWorks, Naive Bayes, 2012) or Support Vector Machines (MathWorks, Support Vector Machine, 2012) etc., the final implementation (which will comprise all the three stages) performed in JAVA.

Acknowledgements
One part of performing this research, the fourth author was supported by the POSDRU/89/1.5/S/63663 grant.

References

Abu-Mostafa, Y.S. (2012). *Learning From Data.* Retrieved Juny 2, 2012, from Learning From Data: http://www.amlbook.com/slides/iTunesU_Lecture18_May_31.pdf

Barbat, B. (2002). *Sisteme inteligente orientate spre agent.* Bucuresti: Editura Academiei Romane.

Blogul lui Ponta. (n.d.). Retrieved July 9, 2012, from http://blogponta.wordpress.com

Crin Antonescu | Blogul Presedintelui PNL. (n.d.). Retrieved July 9, 2012, from http://www.crinantonescu.ro/Blog/CrinAntonescu.html

Dictionar de sinonime online. (n.d.). Retrieved November 22, 2011, from http://www.dictionarsinonime.ro

Dictionar explicativ al limbii romane | DEX online. (n.d.). Retrieved November 22, 2011, from http://dexonline.ro/

DMG. (2012). *Data Mining Group*. Retrieved February 3, 2012, from Data Mining Group: http://www.dmg.org/

Draganescu, M. (2000, ianuarie-aprilie 1–2). Constiinta, frontiera a stiintei, frontiera a omenirii. *Revista de filosofie*, pp. 15–22.

Filip, F.G. (2007). Sisteme suport pentru decizii, Editura Tehnica, Bucuresti.

Gifu, D. & Cristea, D. (2011). Computational Techniques in Political Language Processing: AnaDiP-2011. In J. J. Park, *Future Information Technology* (pp. 188–195). Berlin: Springer.

Gifu, D. & Cristea, D. (2012). Multi-Dimensional Analysis of Political Language. In J. J. (Jong Hyuk) Park, *Future Information Technology, Application, and Service* (pp. 213–221). Springer Netherlands.

Gifu, D. (2010). *PhD Thesis (Abstract) Discursul presei scrise si violenta simbolica. Analiza unei campanii electorate.* Retrieved January 10, 2012, from http://www.uaic.ro/uaic/bin/download/Academic/Doctorate_martie_2010/GfuC.Daniela.pdf

Google.com. (n.d.). *Google chart Tools – Google Developers*. Retrieved February 3, 2012, from https://developers.google.com/chart/

MathWorks, Naive Bayes. (2012). *Naive Bayes classifier – MATLAB*. Retrieved Juny 19, 2012, from http://www.mathworks.com/help/toolbox/stats/naivebayesclass.html

MathWorks, Support Vector Machine. (2012). *Train support vector machine classifier – MATLAB*. Retrieved Juny 19, 2012, from http://www.mathworks.com/help/toolbox/bioinfo/ref/svmtrain.html

Hopcroft, J., Motwani, R., Ullman, J. (1979). Introduction on automata theory, languages and computation, Editor Addison-Wesley.

WordNet Search – 3.1. (n.d.). Retrieved October 15, 2011, from WordNet: http://wordnetweb.princeton.edu/perl/webwn?s=data+mining&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=