An Intelligent Method to Process Romanian Language Internet Reviews

Versavia-Maria Ancusa
Computer and Information Technology Department,
Politehnica University of Timisoara,
Piața Victoriei 2, Timișoara 300006, Romania
Phone: 0256 403 000
versavia.ancusa@cs.upt.ro

Olimpia Ban
Economics Department,
Faculty of Economic Sciences, University of Oradea,
Str. Universitatii, 1, Oradea, Bihor, 0, Oradea, Romania
Phone: 0259 408 105
oban@uoradea.ro

Marian Cornea
Computer and Information Technology Department,
Politehnica University of Timisoara,
Piața Victoriei 2, Timișoara 300006, Romania
Phone: 0256 403 000
marian.cornea@student.upt.ro

Abstract

Internet reviews are a valuable information mine, however most research is oriented towards English based ones. The Romanian language reviews exhibit specific grammar rules, dialect challenges and polymorphism, which need customized methods to be dealt with. This paper offers a method for aggregating heterogeneous Romanian language reviews into a homogenous corpus, fit for further analyse.

Keywords: electronic Word of Mouth (eWOM), mesoscopic approach, natural language processing, complex networks, amfostacolo.ro

1. Introduction

Internet reviews can be seen as a modern communication form, adapted to the highly interconnected digital world of today, distributed in a relationship form of 1 to many, encouraged by social conscience, and generally not (direct) monetary gain, therefore representing an electronic word of mouth (eWOM) paradigm (Yoo, Gretzel, & Zach, 2011).

The main difference between this paradigm and the traditional word of mouth paradigm consists in the impersonal communication format (Chung & Buhalis, 2008), with the person writing the review and the one reading it belonging to different social groups, having different status, education and most importantly, having never met one another. In fact, some researchers (Ayeh, Au, & Law, 2013) consider that the differences make the message more convincing as a whole, while niche messages create a more personalized image of the product / service for their targets. This reflects the social network evolution (Barabasi, 2012), in which clusters naturally evolve, leading the individual consumers to trust reviews from persons that resemble their own characteristics, even though they do not actually know them (Litvin, Goldsmith, & Pan, 2008), (Park & Allen, 2013). Even more surprising is the way people evaluate the credibility of the eWOM, not through a rigorous, conscious decision process, but mostly based on peripheral cues (Metzger, Flanagin, & Zwarun, 2003) and subsequent personal details clustering (Park & Allen, 2013). The reliance on this globalized feed-back method has reached the level in which on-line reviews are more credible than old-fashioned information sources (Dickinger, 2011).

The sheer magnitude of the eWOM is reflected, from a Computer Science perspective, in the presence of massive user-generated data, that can be aggregated and analysed using data mining techniques. eWOM can be analysed at an individual (influencer) or market level (Cheung & Thadani, 2010) or from an input-process-output perspective (Chan & Ngai, 2011), but in any case, eWOM is represented as big data, therefore subjective to all the specific analytics stages and challenges.

According to (Fisher, DeLine, Czerwinski, & Drucke, 2012), big data analytics has the following stages: the data acquiring phase, the modelling phase (which can be further split into architecture choosing and data moulding onto that architecture), the coding/debugging phase and, finally, the reflection phase. In itself, data analytics, either big or small, is an "inherently exploratory" (Fisher, DeLine, Czerwinski, & Drucke, 2012) pursuit, which leads to rapid analytics process successions, which, in turn can introduce errors, that propagate swiftly, compromising the entire process. However, most errors in working with big data are introduced during the data sampling, as well as data cleaning stages (Boyd & Crawford, 2011), which appear at the beginning of the analysis process. Practical surveys point that data cleaning tends to happen after a first iteration of the analytics process, when "a model looked odd". (Fisher, DeLine, Czerwinski, & Drucke, 2012)

While data sampling and cleaning have strong roots in the Statistics field (Kachigan, 1986), new challenges arise due to the vastly heterogeneous nature of the data. Classical statistical algorithms are not fit for the current context in which data is evolving, taking new forms, with each form partially complementing the other, enriching the larger picture. Novel methods and methodologies need to be created in order to operate in such shifting circumstances.

The analytics process in itself can be applied either continuously, concurrent, yet succeeding the data collection process, in which case the data can be expanded/refined on the fly, or the analytics process is applied strictly after the data collection, in which case no further fine-tuning is possible. Each analytics type (static / dynamic) has different data cleaning issues: while the static options allows for more complex, time-consuming, refined algorithms to be applied, no further confirmation from the data sources of the correctness in the cleaning process can lead to unnecessary data censoring. On the other hand, dynamic cleaning has obvious time-constraints, but can be more easily improved over time.

Data cleaning can be performed through a direct manipulation interface or through a script (Fisher, DeLine, Czerwinski, & Drucke, 2012). While either solution will get results, the traceability quality implied by the script makes this a preferred option in practice.

In most cases, the results of the analytics process must often be conveyed to an audience that has little to no expertise in the analytics and/or statistics (Fisher, DeLine, Czerwinski, & Drucke, 2012) so an exchanging of ideas in a simple, common and clear manner is essential. An easy way to do this is to represent the results through pictures, as visual representations tend to be the most common ways of internal information processing (Goodale, 2014), (Healey & Enns, 1999). Conversely, this raises another problem i.e., the limitation of information visualization at a few million data points per screen (Shneiderman, 2008). On the other hand, data in itself is not enough, and not every part of it is equally important, interactions are what brings data to life, which indicate a new opportunity for visualization, through complex networks (Barabasi, 2012).

The purpose of this paper is to present a hybrid methodology for building a eWOM data cleaning algorithm, and apply it for the Romanian language. The algorithm will be tested on the review database from amfostacolo.ro with the stated purpose of building a complex network text representation of that database.

2. Text analysis

Many attempts have been made to model various human languages, leading to the emergence of three perspectives: the microscopic "collection of utterances" view, the macroscopic "set of grammar rules and a vocabulary" view and the mesoscopic hybrid "basic units and emergent interactions" view (Choudhury & Mukherjee, 2008) Complex networks, due to their inherent nature of interaction portrayal, are especially fit to represent the mesoscopic view (Mihalcea & Radev, 2011).

Referring strictly to linguistic networks, the two main uses for this particular representation are: (1) the discovery of languages' inherent properties and (2) any type of knowledge manipulation (machine translation, information retrieval, summarization systems, natural language patterns, etc.). While the networks presented in Table 1 are used to depict language in order of ascertaining their properties, in different contexts, using different corpuses, they can constitute the basis of exploratory systems, such as natural language patterns – machine translation, information retrieval and summarization systems. (Choudhury & Mukherjee, 2008)

Table 1: Natural language processing networks

Network Type	**Graph Type**	**Nodes**	**Edges present based on:**	**Use**
Lexical network	Undirected	Words	phonetic and semantic similarity	exponential degree distribution, high clustering coefficient
Collocation network	Undirected	Words	co-occurrences in similar contexts	power-law distribution the presence of a core-lexicon
Syntactic dependency network	Directed	Words / parts-of-speech	grammatical (logical) relation	Disassortative mixing, hierarchical organization, Small world structure
Phonological networks	Undirected	Sub-lexical units (ex: phonemes, syllables)	Co-occurrences in similar contexts	a power law with an exponential cut-off towards the tail distribution high clustering coefficient strong patterns present

In natural language processing, it can be argued that the context is crucial in determining the underlying value of one word (Turney & Pantel, 2010). The distributional hypothesis states that the context is what defines the semantic of one word, therefore, similar contexts containing different words will give associative qualities to the differences.

Context includes the language in which texts are written and need to be analysed. If most of the scientific literature focuses on English language, there are some attempts to customize it for other languages like Chinese, German, Spanish, Turkish, Arabic, Japanese, Polish and even Romanian (Li, Ye, Zhang, & Wang, 2011), (Feraru, Teodorescu, & Zbancioc, 2010). However, most of these works focus on an academic corpus, which provides limited relevance in dealing with eWOM, since the patterns of the written and casually written/spoke language vary greatly. The research that deals with eWOM is not customized for Romanian, therefore presenting the opportunity to craft such an algorithm.

3. Experimental algorithm development

In order to create the algorithm, we decided on a mixed static – dynamic approach, based on a three-phase process. The process starts with the collection and data aggregation, continues to the second stage, the unification or the basic processing, with the last stage focusing on the analysis

(Figure 1). The analysis part is presented in a dedicated paper (Ban, Ancusa, Bogdan, & Tara, 2015) since it involves industry-specific aspects and further handling specific for the data's origin domain. This paper focuses on the data aggregation and cleaning part.

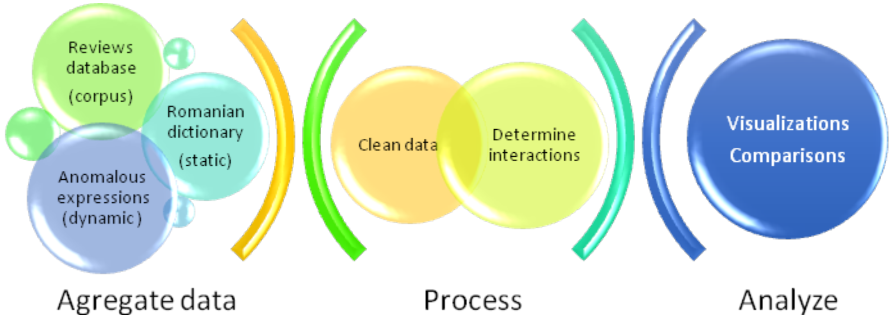

Reviews database (corpus)
Romanian dictionary (static)
Anomalous expressions (dynamic)
Clean data
Determine interactions
Visualizations
Comparisons
Agregate data
Process
Analyze

Figure 1. The eWOM aggregation, processing and analysis

The corpus used was an online review database, consisting of 15200 reviews, written by 8912 different authors with Gaussian age distribution (Figure 2) as to represent a significant viewpoint on the modern Romanian language. The database structure included columns for unique identification of the author, trip geographical details (place, quality, etc.), a written review describing the experience, together with a trip satisfaction numerical score.

The reviews written quality varied greatly from very correct to very colloquial, almost to the point of losing readability. A first aggregation decision was to include even very shoddy and philistine expressions as part of the corpus, together with the cultured expression of the language, because of two factors: (1) the evolutionary aspect of the language manifesting itself and (2) further sentiment analysis must include all forms and nuances because they are relevant in that context.

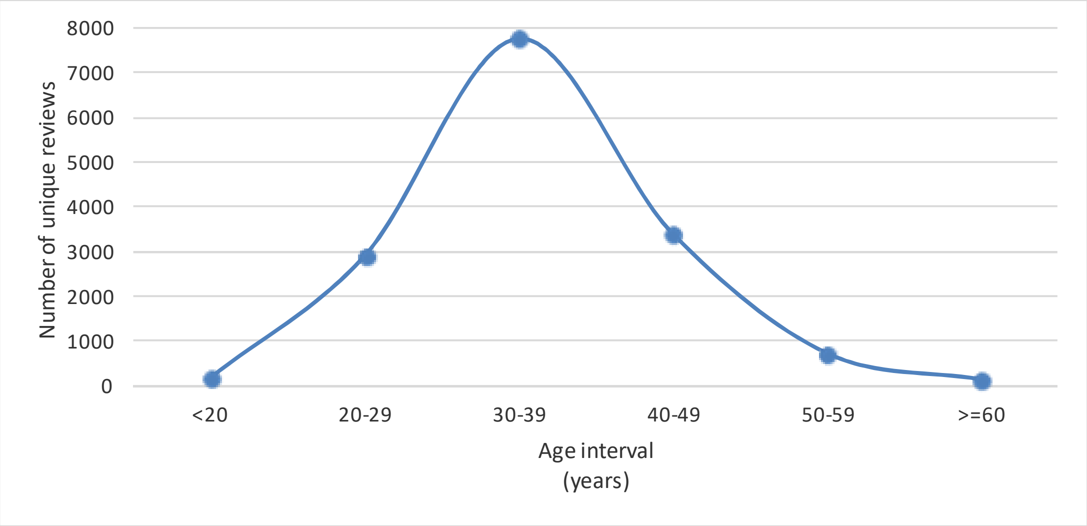


Figure 2 Age distribution of the reviews

Next step (partially based on (Fisher, DeLine, Czerwinski, & Drucke, 2012)) consisted in assigning meaning to the data values. Upon further analysis of the database we discovered four data categories:
1 – dictionary form words (e.g.: “am mers”, “minunat”, “mieunat”)
2 – words with spelling mistakes (e.g.: “am mrs”, “mniunat”, “mienuat”)

3 – common alliterations, new words, colloquial (e.g.: “merem”, “miunat”, “mai”, “pt”)
4 – connector words, exclamations (e.g.: “wow”, “oooooaaaa”, “cu”)
5 – numerical values, punctuation (e.g.: “10”, “!!!!!”, “?!”)

In order to achieve corpus consistency, each word was tested using two databases. The first database used was the Romanian language dictionary, where if a match was found, the word was reduced to its basic declination (e.g.: “minunatul” → “minunat”, “mergem” → “merge”). While this measure clearly affects the eligibility of the written text, it is non-important for a linguistic network in which patterns and centrality matter more than logical text contiguity.

Next, we had to build a special database for the non-dictionary words. This was built dynamically, as our analysis progressed. Each non-dictionary word was added to the database together with a processing directive. The categories developed are presented in Table 2.

Table 2. Categories developed in the dynamic correction database

Category	Description	Action	Example
1	clear spelling mistakes in which the word was easily recognizable	replace misspelled word with dictionary form one, for all further occurrences, prompt only once for solution	“măncărică” → “mâncare”
2	spelling mistakes in which the word is not easily recognizable	prompt supervisor for replacement option, show context *	“miunat” → “mieunat” *or* “minunat”
3	common alliterations	replace word with dictionary form one, for all further occurrences, prompt only once for solution	“pt” → “pentru”
4	colloquial occurrences	Prompt user for action: replace with user-provided solution or delete for all further occurrences	“măi” → *delete* “buuuuun” → “bun”
5	foreign language imports	Prompt user for action: replace with user-provided solution, delete once or delete for all further occurrences	“omg” → *delete for all* “pls” → *delete once* “pls” → “rog”
6	Web-specific elements	delete for all further occurrences *	“http://photos.adress” → *delete for all*

* Note: (1) This action could be further developed into a neural network with a training set consisting of this database. (2) To this date this action allows the use of wildcards like * and ?

Immediately after this phase we proceeded to remove inconsequential items: connector words (prepositions, auxiliary verbs), numerical values, exclamations and punctuation signs. Some of these items might have been removed at the previous step, but to make sure, on the dynamic database we added all these options, starting from their dictionary form, with the associated action *delete for all*.

Resuming, the data values meaning is reflected in four main categories, similar with the model presented in (Fisher, DeLine, Czerwinski, & Drucke, 2012), except for the missing values case:

1. Already “clean” data: dictionary form words
2. Corrupted data: common spelling mistakes

3. Evolved data: common alliterations, foreign language imports, colloquial occurrences
4. Ignored data: numbers, connectors, exclamations, punctuation, hyperlinks

By ignoring punctuation we took a calculated risk, since punctuation can change the meaning of an expression. The reason behind our decision was that in the analytics part of the process, associations were more important than connotations and one negative association due to sarcasm or irony could not overstate the regular ones.

It is worth mentioning that the whole process suffered a second and a third iteration, when during the analytics phase some unexpected results emerged. The common thread behind these occurrences was the presence of unexpected patterns that passed the previously described filters. For example, the occurrence "om mers" was interpreted as clean data since each word on its own represents a correct dictionary input. However, in this case, the regional aspect of the language intervened, by substituting "am" with "om". While the other iterations were based on dictionary and grammar based rules, as mentioned in Table 2, automating the cleaning process, this regional aspect required special handling. Therefore, a new set of rules were manually created to handle such polymorphism occurrences.

During the third iteration, similar word forms, representing different notions were tackled. Such is the case of the noun "zi" with the verb "zice", in which particular case the verb has a form similar with the noun. In order to solve this problem, we implemented, in a case-by-case manner, solutions. In this particular case, the noun was replaced with its articulated form "ziua", requiring user prompting in order to vet the replacement. This solution needs strong further automatic handling, as it is otherwise very time-consuming and error-prone. The rules from Table 2 are obviously incomplete due to the nature of the Romanian eWOM and every iteration required human intervention in the creation of new rules for solving specific problems. By no means have we considered the final result correct, however, it is enough for a complex network mesoscopic analysis that allows errors as long as they are not in a too high number.

After data cleaning, but before analysis, the final step is to create the co-occurrence complex network. The original network, without any processing, had 15.397.933 nodes, or singular occurrences. Through the data cleaning process 4.011.135 deletions were made. After substitution and dictionary form reduction this left only 82.422 unique nodes, a definite improvement from the original size (5.33%).

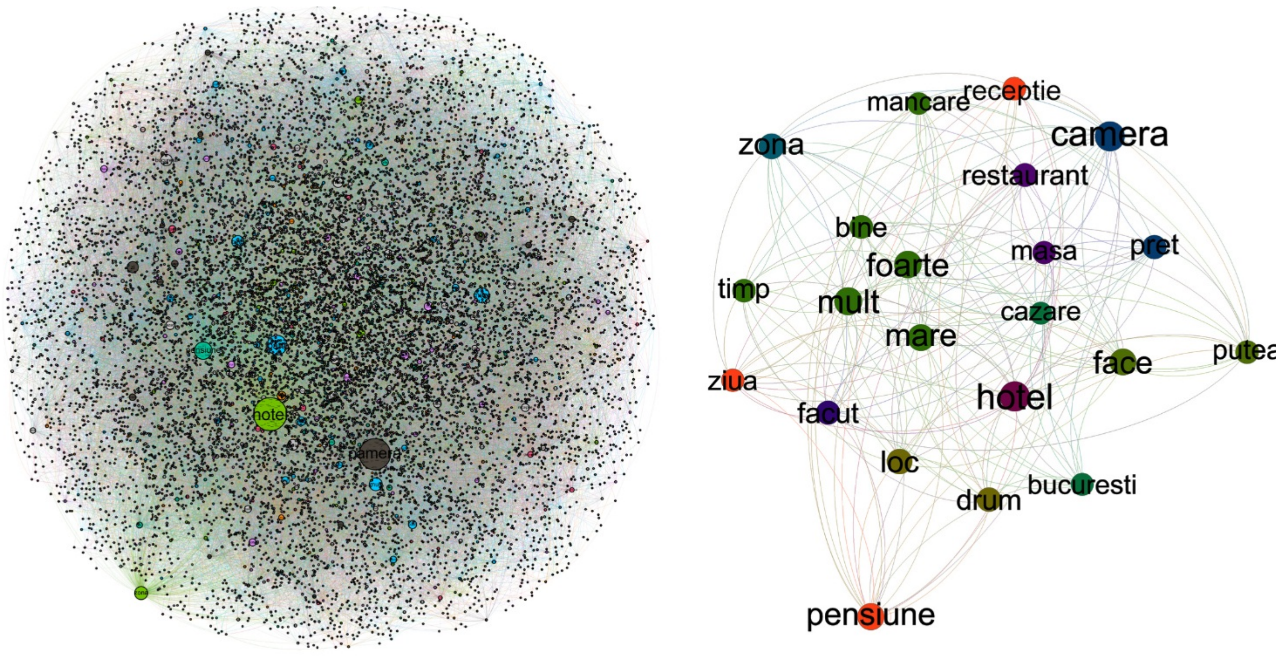


Figure 3. Data as a network – before (left) and after (right) data cleaning

Figure 3 depicts part of the network before and after data cleaning, selected for legibility. The reduction is quite significant, even more poignant due to the graph density factor. As measurement, the initial average for each node was approximately 95 edges, rendering the visual analysis of the human researcher (almost) useless, while the final network is a lot "cleaner' and easier to work with.

6. Conclusion and future work

This paper presented a method used to prepare for analysis big data Romanian eWOM, taking into account language specific aspects. While the method was further used to gather insights, it is not without downfalls: it relies very much on human input and decisions, it needs to be constantly updated to keep pace with language evolution. Further research and work will focus on automating these stages, using machine learning algorithms that detect unexpected patterns and determine new rules for them, simplifying the researchers' work.

Acknowledgment:

We would like to express our gratitude for the support in this research to Mr Cornel Bociort, the site developer of amfostacolo.ro.

References

Ayeh, J., Au, N., & Law, R. (2013). Do We Believe in TripAdvisor? Examining Credibility Perceptions and Online Travelers' Attitude toward Using User-Generated Content. *Journal of Travel Research,* 4(52), 437-452.

Ban, O., Ancusa, V., Bogdan, V., & Tara, I. G. (2015). Empirical Social Research to Identify Clusters of Characteristics that Underlie the Online Evaluation of Accommodation Services. *Revista de Cercetare si Interventie Sociala,* III(50), 293-308.

Barabasi, A. L. (2012). The network takeover. *Nature Physics,* 8, 14-16.

Boyd, D. & Crawford, K. (2011). Six Provocations for Big Data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.*

Chan, Y. Y. & Ngai, E. (2011). Conceptualising electronic word of mouth activity: An input-process-output perspective. *Marketing Intelligence & Planning,* 29(5), 488 - 516. Retrieved from http://dx.doi.org/10.1108/02634501111153692

Cheung, C. M. & Thadani, D. R. (2010). The Effectiveness of Electronic Word-of-Mouth Communication: A Literature Analysis. *23rd Bled eConference eTrust: Implications for the Individual, Enterprises and Society.* Bled, Slovenia.

Choudhury, M. & Mukherjee, A. (2008). The Structure and Dynamics of Linguistic Networks. In *Dynamics on and of Complex Networks: Applications to Biology, Computer Science, Economics and the Social Sciences* (pp. 145-166). Boston, USA: Springer.

Chung, J. Y. & Buhalis, D. (2008). Web 2.0: A Study of Online Travel Community. (Springer, Ed.) *Information and Communication Technologies in Tourism*, 70-81.

Dickinger, A. (2011). The Trustworthiness of Online Channels for Experience and Goal-Directed Search Tasks. *Journal of Travel Research,* 4(50), 378-391.

Feraru, S. M., Teodorescu, H. N., & Zbancioc, M. D. (2010). SRoL - Web-based Resources for Languages and Language Technology e-Learning. *Int. J. of Computers, Communications & Control, V*(3), 301-313.

Fisher, D., DeLine, R., Czerwinski, M., & Drucke, S. (2012, May - June). Interactions with Big Data Analytics. *Interactions*, 50-59.

Goodale, M. A. (2014). How (and why) the visual control of action differs from visual perception. *Proc Biol Sci, 281*(1785). doi:10.1098/rspb.2014.0337.

Healey, C. G. & Enns, J. T. (1999). Large Datasets at a Glance: Combining Textures and Colors in Scientific Visualization. *IEEE Transactions on Visualization and Computer Graphics,* 5(2).

Kachigan, S. K. (1986). *Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods.* New York: Radius Press.

Li, Y., Ye, Q., Zhang, Z., & Wang, T. (2011). Snippet-Based Unsupervised Approach For Sentiment Classification Of Chinese Online Reviews. *International Journal of Information Technology & Decision Making,* 10, 1097-1110.

Litvin, S., Goldsmith, R., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management,* 29, 458-468.

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003). College student web use, perceptions of information credibility, and verification behavior. *Computers & Education* (41), 271-290.

Mihalcea, R. & Radev, D. (2011). *Graph-based Natural Language Processing and Information Retrieval.* Cambridge: Cambridge University Press.

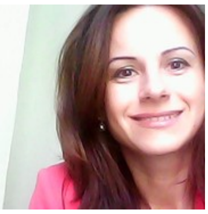
Park, S. & Allen, J. (2013). Responding to Online Reviews: Problem Solving and Engagement in Hotels. *Cornell Hospitality Quarterly,* 54(1), 64-73.

Shneiderman, B. (2008). Extreme visualization:Squeezing a billion datapoints into a million pixels. *Proc. of the ACM SIGMOD International Conference on Management of Data* (pp. 3-12). New York.

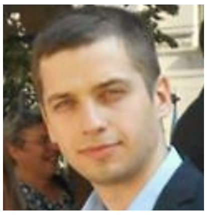
Turney, P. D. & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research,* 37, 141-188.

Yoo, K., Gretzel, U., & Zach, F. (2011). Travel Opinion Leaders and Seekers. *Information and Communication Technologies in Tourism: Proceedings of the International Conference* (pp. 525-535). New York: Springer.

Versavia-Maria ANCUSA (b. March 11, 1981) received her BSc in Computer Science (2004), MSc in Advanced Computer Systems (2005), and PhD in Computer Science (2009) from “Politehnica” University of Timisoara. Now she is a Senior Lecturer in Department of Computers and Information Technology, Automation and Computer Faculty, “Politehnica” University of Timisoara. Her research crosses several domains, including Computer Science, Network Science, Medicine, Marketing and Linguistics.



Olimpia BAN (b. February 23, 1978) received her Bachelor Degree of Finance from University of Economic Sciences Oradea (1997) and PhD in Economics (2005) from the West University of Timisoara. She is now a Professor at University of Oradea, Faculty of Economic Sciences, Department of Economics. Her main research interests are Tourism and Marketing, especially in a national context, with educational and practical applications.



Marian-Gelu CORNEA (b. May 11, 1991) received his BSc in Computer Science (2013) from “Politehnica” University of Timisoara. He is now pursuing his MSc in Advanced Computer Systems at the same university, while working for Autoliv in program development. His main research interests are Python Programming and Agile Development.