Memoirs of the Scientific Sections of the Romanian Academy
Tome XXXIV, 2011

COMPUTER SCIENCE

INTERNET SURVEILLANCE: CONSUMERS OPINION ABOUT CERTAIN PRODUCTS OR EVENTS[1]

ADRIAN IFTENE, ALINA-ELENA MIHĂILĂ, GEORGE-ALEXANDRU VLAD
and GETA STANCU

"Al. I. Cuza" University of Iasi, Faculty of Computer Science
"Al. I. Cuza" University, Faculty of Computer Science,
General Berthelot Street, No. 16, Code 700483, Iasi, Romania
E-mail: {elena.mihaila, george.vlad, geta.stancu}@infoiasi.ro
Corresponding author: adiftene@infoiasi.ro

This paper approaches and tries to offer a solution to a modern day concern: relevant information retrieval and surveillance through the internet. The increasing volume of information that exists on web sites, forums or personal web pages, makes the process of searching for information that is relevant for us to become very complex and time consuming. In addition to their success, social networks like Twitter, MySpace, Facebook, Flickr have more and more users with common interests, and their gathered information has an increasing volume. In such social networks starting from a topic, users can freely express their opinions, add links or relevant photos. The system that we have built uses information collected from the Internet and offers users an easier way to find out positive and negative opinions about a topic. Information is searched on web pages (we prefer blogs, forums or users comments) and on Twitter. Identifying and classifying opinions is done by identifying some emotional triggers and by calculating some valences related to the context in which they appear.

Key words: Internet Surveillance, Information Retrieval, User's opinions and sentiments.

1. INTRODUCTION

In the past few years, new user communities were built on social networks, forums, or sites created for specific classes of persons on the Internet. In this communities, users freely express their opinions about common topics, critic or approve certain aspects related to their common topic. What is interesting is the fact that, for example, in case of a product, based on user opinions, we can have an overview of the product quality; we can identify the advantages of using this product and even its weaknesses.

For example, the web site Tweetfeel[2] offers you the possibility to search on Twitter[3] for the latest tweets regarding a specific topic. These tweets are classified in

[1] This paper is an extended version of paper presented at the conference ConsILR2010.

two categories: negative or positive, based on keywords that are found in users tweets. Classification is done using the distance between keywords that describe the categories, and the words that we are searching for on Twitter, without taking into account the context in which these keywords appear. Therefore, the classification of the tweets is not always correct. Nevertheless, the contexts in which the keywords are used by users in searching are helpful for generating an overview of the topic.

This paper focuses on some concepts related to surveillance of users and information. The purpose of our application is to find out user opinions regarding certain products or events. In the second part we present two case studies to show how our application works: in the first one, we will find out how we can get users opinions regarding certain customs about Easter and in the second one, the difference between the opinions of Romanians and Americans about giving the Nobel Prize to American President Barrack Obama.

2. SURVEILLANCE OF USERS AND INFORMATION OVER THE INTERNET

Foucault considers that society acts like surveillance and a disciplinary society. In this society "*the individual is carefully fabricated in it, according to a whole technique of forces and bodies*" [2].

For Giddens, surveillance means the accumulation of information defined as symbolic materials that can be stored by an agency or community as well as the supervision of the activities of subordinates by their superiors within any community [4]. The modern nation state was from its beginning an information society, because it collects and stores information about citizens (births, marriages, deaths, demographic and fiscal statistics, 'moral statistics' relating to suicide, divorce, delinquency, etc.) in order to organize the administration.

[2] Tweetfeel: http://www.tweetfeel.com/
[3] Twitter: http://twitter.com/

For Fuchs, Internet surveillance is related to information surveillance over the Internet and it is done for different reasons than those related to national security (especially after the 11th of September 2001 terrorist attack) and marketing interest for big companies [3].

Most techniques that imply Internet surveillance involve monitoring data and Internet traffic. Computers connected to the Internet communicate via messages, which are split into small pieces called packages. These packages are then sent through a computer network, node-by-node, until they reaches their destination. At the destination, all the packages are reassembled, reconstructing the initial message. Following the trace of these packages and their content is also related to the Internet surveillance.

In fact, the concept of internet surveillance stands in direct opposition to the idea of a "secure" Internet. According to Netlingo[4], "*Information travelling on the Internet usually takes a circuitous route to its destination computer, through several intermediary computers. The actual route is not under your control. As your information travels, each intermediary computer presents the risk that someone will eavesdrop and make copies. An intermediary computer could even deceive you and exchange information with you by misrepresenting itself as your intended destination. These possibilities make the transfer of confidential information, such as passwords or credit card numbers, susceptible to abuse*."

In order to identify users' opinion on certain products, we have created several modules that perform Internet surveillance, aiming at extracting new useful information for our application. Therefore, we considered the results obtained after searching the Internet using Google search engine, filtered to forums, blogs. Similarly, we perform a search on social networks using Twitter API.

3. SYSTEM PRESENTATION

[4] Netlingo: http://www.netlingo.com/

The system we have built uses two ways of extracting desired information from the internet (See Figure 1). The first consists of a simple interface in which the user can insert a query. This is then processed using the question processing tool presented in [7] and from it we extract the key words after which we proceed with the internet search. This interface uses the libraries provided by the Google API[5] (Google AJAX Search API[6]) to extract the links that are most relevant to our search. Afterwards, the system calls a Lucene Nutch[7] component which saves and indexes the content of these links locally. In the end, also using the Nutch component, we search the index for user reviews, which we classify into positive or negative comments using a special module used to identify opinions and feelings in texts.

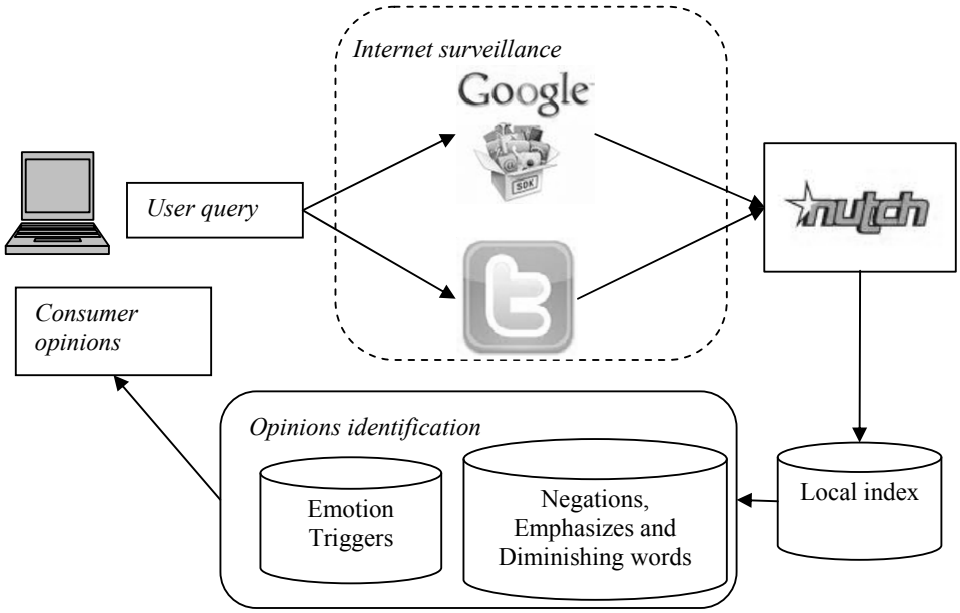

Fig. 1: System architecture

The second looks a lot like the first one, with the difference that instead of using the Google Search API, it uses the API provided by Twitter[8].

[5] Google API: http://code.google.com/
[6] Google AJAX Search API: http://code.google.com/apis/ajaxsearch/
[7] Lucene Nutch: http://lucene.apache.org/nutch/
[8] Twitter API: http://apiwiki.twitter.com/

Next, we present the main components of the system, and their main characteristics.

3.1 GOOGLE AJAX SEARCH API

Search engines are becoming an increasingly important part of our day to day lives. Large volume of information on the Internet makes it impossible to search data without the use of such instruments. We make use of such modules in our application to increase efficiency, and to better interact with the information on the Internet.

In order to incorporate the Google search Engine we use the JavaScript libraries from Google AJAX Search. These allow us to include a control object used to search the Internet for the desired information.

From the search made by the user, Google Search API returns a list of links towards the most relevant web pages. These pages are then copied locally and indexed using the Nutch component.

3.2 APACHE NUTCH

Nutch/Lucene is an open-source platform created in Java. The Nutch component is used for copying and indexing websites. After the indexing process, Nutch allows us to perform searches within the created index. The extracted information is saved as a set of documents known as the Corpus. In the next stage, this Corpus is indexed and made ready for user query. As with most search engines, the searching component is the most complicated, but also it is the most important one. The user query is transformed into a set of index terms, which are send to the Nutch query engine. The result consists of the most relevant documents, i.e. the documents in which the most matches with the user query are found.

3.3 TWITTER API

The Twitter API is based entirely on the HTTP protocol. The methods for retrieving data from the Twitter API require a GET type request, and the methods that insert, modify or destroy data require a POST type request. DELETE requests are also allowed when destroying some information.

The Twitter API allows the request extension format to be modified in order to obtain results in other formats. From the formats currently supported by this API (XML, JSS, RSS or Atom), our application uses the XML format. From the functions provided by this API, we have used mainly those that allow us to extract the desired information from user posts.

3.4 IDENTIFYING POSITIVE OR NEGATIVE USER OPINIONS

Identification of words meanings in texts has recently become one of the main research topics in computational linguistic. Few relevant papers are [5], [9] and [10].

In order to identify user's opinions, we have used a method similar to that which is used in [6]. The method incrementally builds a lexical database (which contains emotion triggering words) similar to that in [1]. The elements in this database allow us to discover emotions in texts extracted by the Google and Twitter APIs. By calculating the global valences of the texts, we can classify the opinion as being positive, negative or neutral.

The base element which identifies emotions is called an "*emotion trigger*" and represents a word or concept which can provide an emotional interpretation of the text content. Here are a few examples of Romanian emotion triggers: "mândrie" (En: "*pride*"), "libertate" (En: "*freedom*"), "stimă" (En: "*esteem*"), "familie" (En: "*family*").

The lexical database has been constructed from 30 terms which are present in "Maslow's pyramid" [8], which have been translated into Romanian. Additionally, we used the Romanian WordNet [11] to extract synonyms, antonyms, and hyponyms for those terms.

After building the lexical term database, the next step was to correlate these terms with certain valences and emotions. For this we take into account the following rules:

- A positive value is given to the main emotion triggering terms and their synonyms.
- Also, a positive value is given to hyponym terms and terms which are derived from the main terms.
- A negative value is given to the terms antonym from those above.
- The valence of any term modifies according with the modifiers the come with them (terms that negate, emphasize or diminish a value).

3.4.1 VALENCE MODIFIERS

In order to determine the final valences of the emotion triggers a set of "*valence modifiers*" is defined [1]. A *valence modifier* consists of a term that modifies the valence of any term it is associated with. The modifiers we have identified are of three types:

- **Negations words** – that modify the valence radically from the positive pole to the negative one or backwards. Here, modifiers are represented by terms which insert negation "*nu*" (En: "*no*"), "*niciodată*" (En: "*never*").
- **Emphasizes words** – that emphasize the positive or negative aspect of a trigger. This set contains adjectives like "*mare*" (En: "*big*"), "*mai mult*" (En: "*bigger*"), "*mai bine*" (En: "*better*"), "*profund*" (En: "*profound*"), "*excepțional*" (En: "*great*") or adverbs that emphasize the understanding of the whole context they are part of, like "*cu siguranță*" (En: "*surely*"), "*sigur*" (En: "*sure*"), "*cert*" (En: "*certainly*"), "*în definitiv*" (En: "*ultimately*").
- **Diminishing words** – that diminish the positive or negative aspect of a trigger, taking it towards a neutral valence. Diminishing words are represented by adjectives like "little", "less", "worse", "rather", by modal verbs, "to be able", "to be possible", "to need", "to want", by adverbs like "possible", "probable". Modal verbs bring in the notions of uncertainty and possibility and diminish the valence and emotion of the whole context they are part of. Also, these diminishing words help us distinguish between events that took place, could have taken place, occur now, or will take place in the future.

For example, how would a passage sound with different valence modifiers: "*este minunat*" (En: "*it is wonderful*") (initial form), "*nu este minunat*" (En: "*it is not wonderful*") (the negative form), "*ar putea fi minunat*" (En: "*it could be wonderful*") (the diminished form), "*e absolute minunat*" (En: "*it is absolutely wonderful*") (the emphasized form).

There are some terms that when put together create the feeling of irony. This is also the case for the next association: "*The exceptional organizer failed to resolve the problem*", where the term, "*exceptional*" (which usually has a positive valence), next to the term "*has failed*" (which usually has a negative valence), conjures that "*The organizer which was so well-known for his skill, failed a simple task*".

Moreover, while testing and verifying the application we have discovered new terms which are specific to discussions on forums, blogs or social networks. With their help, we have completed the resources obtained in the previous steps. Here are some examples of specific terms taken from English "bravo", "super", "fine", "good" or successions of special characters that represent emotional icons: ":)" (Smiley face ☺), ":(" (Sad face ☹), etc. An important observation from [1] that needs to be mentioned here is the following: in order for a valence modifier to fulfill its purpose (to modify the valence of the terms), it needs an attitude to have been mentioned in the text (whose understanding can be modified). For example, in the sentence "*John is home."* that only presents a fact and not an attitude, introducing a negation, "*John is not home*", doesn't change the valence of any term.

4. CASE STUDIES

In this chapter we have conducted two case studies, in order to show the working mode of our system. With the help of this system, users will have all this information centralized and moreover it will be classified into positive or negative comments. Furthermore, given the fact that the application accesses multiple internet areas and offers the information in a centralized manner, helps users reduce the time needed in order to obtain such information.

4.1. THE EASTER CELEBRATION

The first case study is related to the Easter celebration. The purpose of the application is to help those who want to create products specific to this event, without having specific networks or time available in order to search them. For this case study we have in mind traditional products ("*cozonac*" (En: "*sponge cake*"), "*pască*" (En: "*bread*") and "*ouă roşii*" (En: "*red eggs*")) and we have identified user opinions regarding certain recipes and preparation methods. The goal was to identify whether we can find, using our application, a recipe that is better suited to the users tastes and the advantages and disadvantages of that recipe.

As stated previously, first, we have searched using the Google search engine and the API associated with it in order to extract this information from simple web pages, forums, blogs and then we have used the Twitter API for extracting such information from posts existing in this social network.

When working with the Google API, to ensure that the results are relevant, we have constructed several queries choosing various combinations that use words which represent either products: *sponge cake*, *bread* or *red eggs*, either ingredients: *ciocolată* (En: *chocolate*), *brânză* (En: *cheese*), *smântână* (En: *milk cream*), *nuci* (En: *nuts*) (with or without diacritics) and moreover we have asked that this information is found on pages that respect the forum or blog format types.

For the Twitter API we have built less complex queries (with less words) in order to have higher rate of success. In the end we have conducted 20 Google searches and 7 searches on Twitter.

For each Google search, Nutch took the first 10 links from the given results and saved their content on the local machine. We have processed their content using the emotion identifying component and we have extracted positive and negative feelings.

Next, we will see a couple of examples of sentences extracted from forums using the Google API. After searching by terms like sponge cake recipe in user comments we have found emotion triggers: positive "se **pare** ca cozonacul mamei a trecut cu **brio** proba **:)**" (En: "it **seems** that mother's sponge cake **passed** the mark with congratulations **:)**"), "Aluatul de cozonac e **foarte delicat** si **sensibil**." (En: "Sponge

cake dough is **very delicate** and **sensitive**."), "E o reţetă **super** simplă şi **super gustoasă**...." (En: "It is a **super easy** and **super tasty** recipe..."), and with a negative inclination "oricât aş vrea **nu reuşesc** sa fac acest cozonac" (En: "as much as I would like, I just can't make this sponge cake"), etc.

Although the Twitter searches were much simpler, the results were considerably less than those returned by the Google engine. From what we have seen, this is because the number of comments written in the Romanian language is very small and many times comments in Romanian contain words or fragments written in English.

Here are a couple of posts obtained from Twitter after searching sponge cake, and red eggs: "deza@zosz_ro: **ciudată combinaţie**...dar suna **gustos**...acum mănânc din primul meu cozonac...si **ciudat**, a ieşit **bun** (En: weird combination... although it sounds tasty, I'm now eating from my first sponge cake and....strangely, it came out good"), "richieTM: **Deci cea mai buna** plăcinta cu brânza e Pasca!! (En: **Definitely** the **best** cheese pie is the bread!!!"), "lau_anca @RaduCeuca **super**! **dar** ai uitat de clasicele oua roşii **:)** (En: **great**! **but** you have forgot about the classic red eggs **:)**"), etc.

From this study we have concluded that the user's opinions on forums and blogs are more consistent and of course more relevant than those given by the users on Twitter which are shorter and many times use images loaded by them.

4.2 AWARDING AMERICAN PRESIDENT BARACK OBAMA WITH THE NOBEL PRIZE

One important controversy in 2009 was whether American President Barack Obama deserved the Nobel Peace prize. In order to better understand the issue we conducted a survey which analyzes the reaction of people from around the world. The data was collected from a series of sites, blogs and forums which expressed different views on the subject.

The obtained results showed that there is a clear difference between the views of the American people (which is overall positive) and that of the Romanian people (which is overall negative). This can be explained by the fact that the majority of positive views came from people who voted for him in the 2008 presidential elections, while the

majority of neutral or negative views came from people who were not directly influenced by Barack Obama's actions.

The application has extracted information from 20 sites (ten from Romania and ten from United States) using Google API, summing up to a total of 327 posts. To identify the "valence" of the fragments extracted from the English resources we translated them using Google Translate[9] and extended them using English WordNet[10].

The 20 sites were obtained as a result of using queries including keywords related to the following subjects:

1. *Ce părere aveţi despre înmânarea premiului Nobel lui Barack Obama*? (En: *What do you think about giving the Nobel Prize to Barack Obama*?)
2. *Care sunt părerile pro şi contra la acordarea premiului Nobel lui Obama*? (En: *What are the pros and cons for awarding the Nobel Prize to Obama*?)
3. *A meritat Obama premiul Nobel*? (En: *Did Obama deserve the Nobel Prize*?)

Looking at the Romanian sites, some of the people who expressed their opinion pointed out that, because of this uninspired decision, the integrity of the Nobel Prize award will suffer from now on. Once this decision was made, most of them believe that the awards lost some of its importance and that from now on winning such an award will not be so significant.

Given below are some examples of posts in which we detected a big emotional charge which corresponds to disapproval and indignation (all had one source: http://economie.hotnews.ro/stiri-media_publicitate-6262253-presa-americana-radiosul-obama-accepta-premiul-pentru-pace-sau-cum-castigi-nobelul-12-zile.htm):

- "…e **culmea tupeului** sa dai premiul Nobel cuiva care era presedinte de doar 12 zile la data inchiderii perioadei de candidatura !!...", (En: "… it is **outrageous** to award the Nobel prize to someone who was elected president just 12 days before the votes were closed !!!"),

[9] Google Translate: http://translate.google.com/
[10] WordNet: http://wordnetweb.princeton.edu/perl/webwn

- "acest Nobel al lui Obama va ramane in istorie ca fiind **probabil cel mai nemeritat** din cate s-au acordat!" (En: "this Nobel prize of Obama will remain in history as **probably** being the **most undeserved** prize ever !"),
- "**Cred** ca prin aceasta **trista alegere**, **nemeritata**, Nobelul pt Pace e **definitiv compromis**." , "a devenit evident ca **nu** are vreo legatura cu **meritele** sau **rezultate** reale obtinute. ". (En: "I **believe** that through this **sad decision**, **undeserved**, the Nobel prize is forever **compromised**", "it has become obvious that it has **nothing to do** with the **merits** and the **results** obtained").
- "**Cred** ca e o **jignire** la adresa celor care chiar fac ceva pentru **pace** in lume...", "....pentru ce?!?! Pentru cele doua **razboaie** in care SUA sunt implicate?" (En: "I **think** it is an **insult** to those who really have one word to say in world **peace** ...", ".... and for what?!?! For the two **wars** in which the United States were involved in?")
- "Un act de o **nesimtire** si un **dispret incredibil** fata de sute de lideri politici" (En: "An act of **disgrace** and **incredible contempt** towards hundreds of political leaders").

Analyzing the English sites (mostly those from the United States) we noticed a different attitude. Although most of them do not approve with this decision, the criticism is not directly pointed towards Obama, but mostly towards the committee who awarded this distinction.

Regarding the emotional triggers, the sites from Romania contain a lot more words with negative "valence" and modifiers that enforce the "valence" (as *incredibil* (En: *incredible*), *total* (En: *totally*), *foarte* (En: *very*)).

In the sites from America we noticed a tendency of using words that neutralize the valences (maybe, possible, probable).

The majority of opinions address a general confusion ("I **don't understand what**", "I **don't see why**", "I **don't know why** he was elected", "I **don't think** he ...", "I **don't consider** that it was the best decision") regarding the decision made.

There were a few people who didn't have a negative opinion about the decision, although they aren't in total agreement with the result. They expressed a good opinion +about the person in question.

5. CONCLUSIONS

This paper presents the main components of the system we offer to users who wish to find negative or positive opinions regarding some products or events. The system is an alternative to the classic search on the Internet, bringing new possibilities of combining the results obtained through a simple search with those extracted from the user's blogs, forums or ever social networks. The main components of the system are based on the components which improve search and extract relevant information via the Internet, namely the search API from Google and Twitter.

The system combines the valences associated with the sentences, the distance between key words and emotional triggers, taking into account the valence modifiers. From our evaluations on 6 different topics of search, with the help of 3 persons that evaluated over 100 extracted paragraphs, we can say that we are at the beginning of the road and this reflects on the quality of the results, which is modest (around 44%).

The main problems come from the fact that the main emotional triggers that we took into consideration were not referring to the keywords of the search. Another important problem was that, in calculating the associated valences, we preferred short sentences, which were not always as relevant as we would like.

As future extensions, we plan to resolve the mentioned problems, which come from the identification of the semantic roles. Also, we wish to accomplish a more relevant evaluation with around 20 different topics and more than 10 human evaluators.

Acknowledgements. The research presented in this paper was funded by the Sector Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944.

Authors' contributions: Adrian Iftene handled the component which discovers the emotions that appear in texts and calculates some global valences of the texts. He created the whole document structure in which he wrote chapters that make reference to feelings. Alina-Elena Mihailă and George-Alexandru Vlad integrated the Google Search API, Tweeter API and Nutch component in the main system. They extracted data, analyzed results and perform the case study about Easter Celebration. In the end they wrote corresponding chapters in this paper. Similar, Geta Stancu performs and writes the case study case for awarding the Nobel Prize to American President Barack Obama. All authors worked on evaluating the final system.

REFERENCES

1. BALAHUR A., MONTOYO A., *Applying a culture dependent emotion triggers database for text valence and emotion classification*, Procesamiento del Lenguaje Natural, ISSN 1135-5948,2008, **40**, 107-114.

2. FOUCAULT M., *Discipline and punish*, Vintage, New York, 1977.

3**.** FUCHS C., *Social Networking Sites and the Surveillance Society. A Critical Case Study of the Usage of studiVZ, Facebook, and MySpace by Students in Salzburg in the Context of Electronic Surveillance*, Salzburg/Vienna, Austria, 2009; Forschungsgruppe "Unified Theory of Information" - Verein zur Förderung der Integration der Informationswissenschaften, ISBN 978-3-200-01428-2, 2009.

4. GIDDENS A., *A contemporary critique of Historical Materialism*, Vol. 1: *Power, property and the state*, Macmillan, London, 1981.

5. HATZIVASSILOGLOU V., MCKEOWN K. R., *Predicting the semantic orientation of adjectives*, Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, Morristown, NJ, USA, 1997, 174-181.

6. IFTENE A., ROTARU A., *User Profile Modeling in eLearning using Sentiment Extraction from Text*, Research in Computing Science, Special issue: *Natural Language Processing and its Applications*, Vol.46, p. 267-278, Instituto Politecnico Nacional, Centro de Investigacion en Computacion, Mexico, 2010. ISSN: 1870-4069; Poster at 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICling 2010), 21-27 March, Iasi, Romania, 2010.

7. IFTENE A., TRANDABĂŢ D., PISTOL I., MORUZ A.M., HUSARCIUC M., STERPU M., TURLIUC C., *Question Answering on English and Romanian Languages*, Proceedings of the CLEF 2009 Workshop, 30 September - 2 October, Corfu, Greece, 2009.

8. MASLOW A. H., *A Theory of Human Motivation*, Psychological Review, 1943, **50** (*4*), 370-96.

9. MIHALCEA R., BANEA C., WIEBE, J., *Learning Multilingual Subjective Language via Cross-Lingual Projections*, Proceedings of the Association for Computational Linguistics (ACL 2007), Prague, June 2007.

10. TUFIŞ D., *Playing with Word Meanings*, In *From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence* (Lotfi A. Zadeh, Dan Tufiş, Florin Gh. Filip and Ioan Dziţac, eds.), Publishing House of the Romanian Academy., ISBN 978-973-27-1678-6, 2009, 211-223.

11. TUFIŞ D., BARBU E., BARBU MITITELU V., ION R., BOZIANU L., *The Romanian Wordnet*, Romanian Journal of Information Science and Technology, 2004, **7**, *1-2*, 107-124.

Received August 16, 2010