INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(6):694-702, December, 2014.

Detecting Emotions in Comments on Forums

D. Gîfu, M. Cioca

Daniela Gîfu
"Alexandru Ioan Cuza" University of Iaşi
16 General Berthelot St., Iaşi, 700483, România
daniela.gifu@info.uaic.ro

Marius Cioca
"Lucian Blaga" University of Sibiu
10, Victoriei Bd., Sibiu, 550024, România
marius.cioca@ulbsibiu.ro

Abstract: The paper presents one of the most important issues in Natural Language Processing (NLP), emotion identification and classification to implement a computational technology based on existing resources, open-source or freely available for research purposes. Furthermore, we are interested to use it for establishing Gold standards in sentiment analysis area, such as SentiWordNet. In this sense, we propose to recognize and classify the emotions (sentiments) of the public consumer from the written texts which appeared on the various Forums. We analyse the writing style which refers to how consumers construct sentences together when they write comments to indicate their passion about an entity (persons, brand, location, etc.). We present in this paper a method for integrating Romanian lexical resources from emotional perspective, in developing, which can be used in sentiment analysis. This study is intend to help direct beneficiaries (public consumer, marketing managers, PR firms, politicians, investors), but, also, specialists and researchers in the field of natural language processing, linguists, psychologists, sociologists, economists, etc.
Keywords: sentiment analysis, language resources, emotions levels, semantic classes, Forums.

1 Introduction

In our context, emotion in writing refers to how public consumers express a personal opinion of their experience about entities (products, persons, tourism objectives, etc.). When we say public consumer, actually, we say any commentator who is interested in a range of information about a particular entity.The option for such a topic, known as sentiment analysis (SA) or *opinion mining*[1], encountered in texts circulated on different *Forums*, and comes from the need to clarify descriptive consumer behavior, affected by the amount of promotional messages, regardless of their nature and purpose. At the present time, sentiment analysis is one of the most studied natural language processing (NLP) issues.

The hypothesis of this paper is that by observing the emotional orientation of the commentators over time (visible in writing style) on Forums can help us to build a database with information on topics, services, products, etc. for the public interest, which can serve to implement a NLP tool, useful to predict potential consumer needs.

The paper is structured in five sections. After a brief introduction about the importance of this study, the section 2 mentions some important works focused on SA. The section 3 describes

[1]Opinion Mining originates from the Information Retrieval (IR) community, and aims at extracting and processing users' opinions about entities (products, movies, etc.). Sentiment analysis was initially formulated as the NLP task of retrieval of sentiments expressed in texts. Looking closely, these two issues are similar in their own essence and fall under the area of Subjectivity Analysis.

four units of sentiment analysis some of the most coomonly used in SA, and section 4 describes the our tool functionality. The last section highlights conclusions and mentions the future work, one of the projects of NLP-Group@UAIC-FII.

2 State of the art

Nowadays, Forum becomes a long-term instrument that can consolidate the public sphere, Habermas's concept [9] and civil society. In opposite to the instrumental view of *liberalization* of the Internet, the new dimension can be classified as *environmental*. The ubiquity of Forums affects the marketing mechanisms to respond to the challenges imposed by it. If the landscape of communication becomes denser, more complex and more participative, then the *network population* gets increased access to information, achieving multiple opportunities by engaging in public speech and putting in motion collective actions. But, a problem appears. More information, more opinions reflected mostly in writing style. In fact, any difference in writing reflects the heterogeneity in reviewers culture, education, occupation and so on. This heterogeneity can be quantified in sentiments.

The sentiment is the overall emotion towards the subject matter expressed by the reviewer. In general terms, SA consists of extracting opinions from text. It is assimilated as *subjectivity analysis* [2] or *evaluating affection* [1]. SA defines the processing search results from an article, generating a list of attributes product (quality, characteristics, etc.) and aggregating opinions for each of them (e.g. poorly, good). Moreover, SA has been interpreted as including various types of analysis and evaluation [14], [15], [17], [18].

Another important dimension of SA is researching objectivity in a text, finally resulting a text classification into two classes - objective and subjective -, frequently more difficult to undertake than for a polarity one [16]. In 2001, sentiment analysis was the subject of two researches by Das and Chen [3], and Tong [1], concerned on the opinions on the market sales. Out attention is also take up by the classification of the degree of positivity of a text (document, sentence/clause, etc.), consisting in opinion words (e.g. angry, happy). For instance, in elections, we established two classes, positive and negative, each of them with other three subclasses for determining the intensity of sentiment [7]. Moreover, in the sentiment analysis area there are approaches that consider, also, the neutral class (value 0), assigning words with one value from -5 to +5, with two classes more than the first author [8]. This paper describes a method with a shorter scale of values, from -1 to +1, as the authors are interested to discover the sentiment extracted from their comments.

3 Units of sentiment analysis

SA offers organizations the possibility to monitor opinions about products/ services and their reputation (e.g. measuring feedback with statistical software packages SAS - *Statistical Analysis System*, SPSS - *Statistical Package for the Social Sciences* or *Superior Performing Statistical Software*), on various Forums platforms in real time and to act accordingly.

We describe below four lexical units for SA.

3.1. Document as the unit of analysis

It is the simplest form of SA and assumes that the document contains an opinion on one main message expressed by the commentator. We will stop at two approaches of sentiment analysis from the document.

a) *Supervised* the document must be classified in a finite set of classes, the training data are assigned to each class. This is for the simple case, when there are two classes: positive and

negative. Also, a neutral class can be added or a numeric scale can be considered from which the document has to be reported (for instance, SentiWordNet). Esuli and Sebastiani [6] reports three sentiment scores: positivity, negativity and objectivity. The system learns a classification model based on the training data, using an algorithm of classification, such as SVM (Support Vector Machines) or KNN (K-Nearest Neighbors). Then, this classification is used for mapping new documents in their different sentiment classes. Good precision is achieved even when each document is represented as a bag of words [13].

b) *Unsupervised* the document is based on determining the semantic orientation (SO) of specific phrases. If the average SO of these phrases is above a predefined threshold, the document is classified as positive. Otherwise, it is considered negative. For instance, a set of predefined part-of-speech (POS) models can be used to select those sentences [21] approach taken into consideration in this study - or to create an opinion lexicon structured in words and syntagmas used by the first author since 2009.

3.2. Sentence as the unit of analysis

For a more refined analysis of opinions about an entity (organization, product, political actor, etc.) we must move to the sentence level. It is assumed that there is only one opinion (sentiment) in each sentence. To prove it, each sentence is splitted in clauses (a fragment with a predicative verb) and every clause contains only one opinion which we classified it in subjective or objective. Only the subjective clauses will be analyzed. For instance, the approach is based on minimal reductions [19], as the premise is that the neighboring clauses should have the same subjective classification. Then the sentences can be classified as either positive or negative.

3.3. Comparative sentiment analysis

In many cases, users do not offer a direct opinion about a product, preferring instead comparable opinions such as:

Dacia Logan arată ***mult mai bine*** *decât Dacia Solenza*[2].

In this case, the purpose of the sentiment analysis system is to identify opinions of the sentence containing the comparative views, as well as to extract there from the preferred entity. Authors like Jindal and Liu [12] describe this analytical method. Using a relatively small number of words as comparative adverbial adjectives *mai mult, mai puţin, uşoare*[3], superlative adjectives and adverbs *mai, cel puţin, cele mai bune*[4], additional clauses *favoare, mare, preferă, decât, superioară, inferior, numărul unu, împotriva*[5], we can cover 98 % of the comparative opinions.

For these words/groups of words which frequently appear in texts, but with low precision, a classifier[6] can be used to filter phrases that do not contain comparative views. Ding, Liu and Zhang [4] present a simple algorithm for identifying preferred entities relating to the type of comparisons used and the presence of negation.

3.4. Sentiment lexicon

As we have seen so far, the lexicon is the most important resource for the majority of the sentiment analysis techniques. There are three options in order to create a lexicon of sentiments:

a) *manual approaches*, when researchers create a manual lexicon, consisting of a set of words selected from explanatory dictionaries that will be subsequently extended by using existing lexical resources (synonyms and antonyms for enrichment). We have already mentioned WordNet. This process requires a laborious effort, especially that each domain needs its own lexicon. A handy algorithm is proposed by Kamps, J., Marx, M., Mokken, R.J. and de Rijke, M. (2004).

[2]En. - *Dacia Logan looks much better than Dacia Solenza.*

[3]En. - *more, less, easy.*

[4]En. - *more, at least, the best, etc.*

[5]En. - *favour, high, prefer, rather than, superior, inferior, the number one, against.*

[6]For example, Naive Bayes classifier, a statistical method for forms classification and recognition, where each document represents a collection of words and word order is considered irrelevant.

b) *corpus-based approaches*, in which a set of words/phrases extracted from a relatively small corpus is extended by using a large corpus of documents of a single domain.

The main disadvantage of any dictionary-based algorithm (a) is that the acquired lexicon is too general and therefore does not capture the specific features of a particular area. Advanced approaches based a lexicon are reported in Dragut et al. [5].

If we want to create a specific sentiment lexicon, we have to use a corpus-based algorithm. A classical work in this area [10] highlights the concept of sentiments consistency allowing the identification of complex polar adjectives. In other words, a set of linguistic connectors *şi, sau, nici, fie, sau*[7] has been used to find the adjectives that are connected to the adjectives with well-known polarity.

For example: *bărbat puternic şi armonios*[8].

If we admit that puternic is a positive word, we can assume that the word armonios is also positive thanks to the use of the connector *şi*.

4 The tool description

This version of our tool[9] is able to detect and to explain the appreciations about some entities (persons, products, brands, etc.). This tool is based on information like labeling of parts of speech (e.g. the XML example), extracting of interest nominal groups, automatic extracting of entities and anaphoric connections.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<DOCUMENT>
<P ID="1">
<S ID="1">
<W EXTRA="NotInDict" ID="11.1" LEMMA="" MSD="Vmip3s" Mood="indicative"
Number="singular" POS="VERB" Person="third" Tense="present" Type="predicative"
offset="0"></W>
<NP HEADID="11.2" ID="0" ref="0">
<W Case="direct" Gender="masculine" ID="11.2" LEMMA="nimic" MSD="Pz3msr"
Number="singular" POS="PRONOUN" Person="third" Type="negative"
offset="1">Nimic</W>
<W ID="11.3" LEMMA="mai" MSD="Rg" POS="ADVERB" offset="7">mai</W>
<W Case="direct" Definiteness="no" Gender="masculine" ID="11.4" LEMMA="odios"
MSD="Afpmsrn" Number="singular" POS="ADJECTIVE" offset="11">odios</W>
<W ID="11.5" LEMMA="," MSD="COMMA" POS="COMMA" offset="16">,</W>
<W ID="11.6" LEMMA="mai" MSD="Rg" POS="ADVERB" offset="18">mai</W>
<W ID="11.7" LEMMA="oribil" MSD="Rg" POS="ADVERB" offset="22">oribil</W>
<W Case="direct" Definiteness="no" EXTRA="NotInDict" Gender="masculine"
ID="11.8" LEMMA="decat" MSD="Afpmsrn" Number="singular" POS="ADJECTIVE"
offset="29">decât</W>
</NP>
<NP HEADID="11.9" ID="1" ref="1">
<W Case="direct" Definiteness="yes" Gender="masculine" ID="11.9" LEMMA="pantof"
MSD="Ncmpry" Number="plural" POS="NOUN" Type="common" offset="35">pantofii</W>
<NP HEADID="11.10" ID="2" ref="2">
<W Case="direct" Definiteness="no" Gender="masculine" ID="11.10" LEMMA="sport"
```

[7] En. - and, or, not, either.

[8] En - strong and harmonious man.

[9] The version previous of this tool, called EAT (Emotional Analysis Tool), is still in testing phase.

```
MSD="Ncmsrn" Number="singular" POS="NOUN" Type="common" offset="44">sport</W>
<W ID="11.11" LEMMA="cu" MSD="Sp" POS="ADPOSITION" offset="50">cu</W>
<NP HEADID="11.12" ID="3" ref="3">
<W Case="direct" Definiteness="yes" Gender="feminine" ID="11.12"
LEMMA="platform" MSD="Ncfsry" Number="singular" POS="NOUN" Type="common"
offset="53">platforma</W>
</NP>
</NP>
</NP>

</DOCUMENT>
```

Moreover it was developed an important ontology of entities, categories and values. In figure 1 we have the interface of our tool. We describe briefly work methodology:

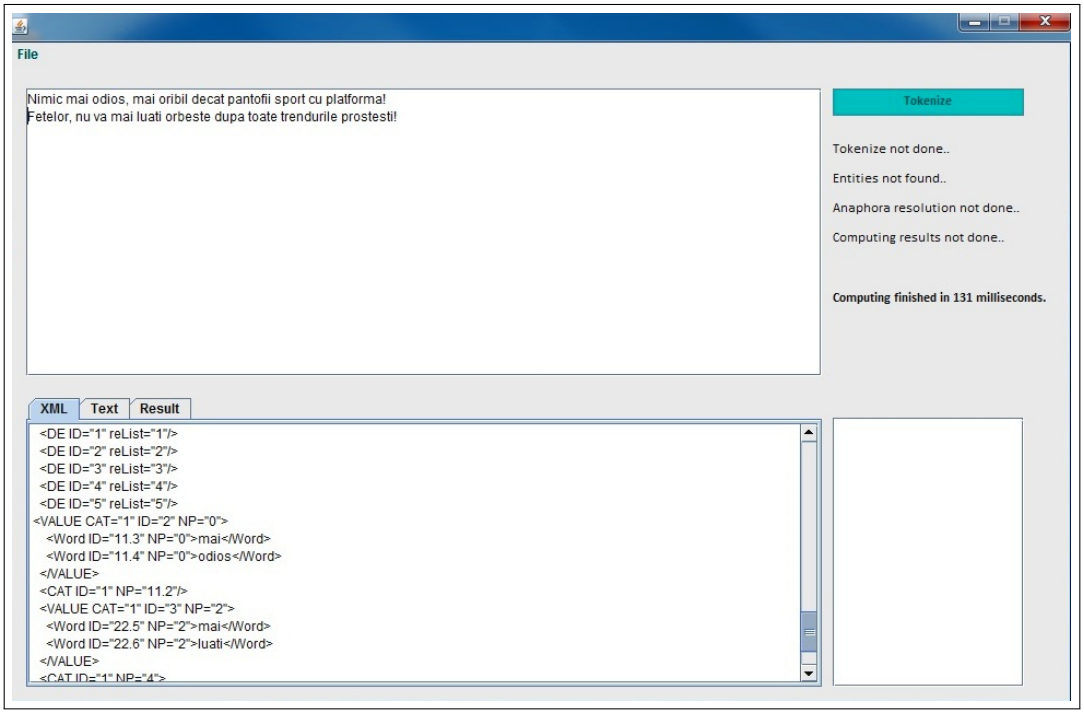

Figure 1: The interface of the computational tool

1. A corpus of texts (50 texts) is manually annotated using PALinka[10], in order to build triplets of the form: `<entitate><categorie><valoare>`.

2. The text is preprocessed using UAIC Romanian Part of Speech Tagger[11] [20]. This tagger combines a statistical model to one based on rules. The morphological dictionary was largely extracted from DexOnline and contains 1.25 milion distinct words. The result is an XML file, each word has been tockenized and annotated according to the POS that it represents.

3. Noun phrases are detected and annotated with NP-chunker[12] [20]. This chunker is used in

[10]http://clg.wlv.ac.uk/trac/palinka/

[11]POS tagger has a precision of 96,6%8, considered on the corrected version of the novel "1984" (George Orwell).(http://instrumente.infoiasi.ro/WebPosRo/).

[12]Chunker receives as input the tokenized text, in XML, formed by suitable groups in text, and the output is another XML file where each nominal interest group will be annotated XML with NP label (http://instrumente.infoiasi.ro/WebPosRo/).

many applications to resolve the ambiguities or to extract information. For example, the newest work studies based on machine translation use texts in two languages (parallel corpora) to derive the appropiate transfer models.

4. Proper names of entities are automatically extracted using a named entity recognizer techonology GATE[13] open source (ANNIE)[14].

5. Anaphoric links (especially, pronouns) are extracted from the text using RARE (*Robust Anaphora Resolution Engine* implemented by Eugen Ignat [11].). This process makes appreciations that the text expresses about those entities (coreferences) to be aggregated to the same entity (reference).

6. Entities, categories and values from the ontologies that have been already created are recognized in the text using NER (Named Entity Recognition) which extracted the entities automatically. NER recognizes entities such as persons, organizations or geographic locations, receiving as input a natural language text and the output is a text file which contains entities as a string that uses separators to delimit named entities.

7. A set of rules is written for the recognition of values and the connections such as `<entity><category><value>` are established.

8. Graphical inteface reveals the extracted information and global scores.

Of the recorded, our tool is able to detect and explain qualitative appreciations about entities. In figure 2 is profiled the architecture of this software as follows:
- *building an anthology* of entities, categories and values, useful to obtain a correct and complete result;
- *preprocessing text*, meaning annotation, splitting text into entities (words, symbols or tokens);
- *noun phrase chuncking* (NP-chunk), meaning splitting text into sequences of syntactically correlated words (nominal groups);
- *recovering anaphoric connections*, important not to lose any reference to a particular entity, using RARE.
- *extracting entities*, using NER module. It receives a file .txt (*input*). The output file contains only the entities mentioned in the analyzed text.

For instance: "[Vodafone] [România] oferă cea mai bună conectivitate pentru serviciile de date dintre toate reţelele mobile [GSM]/ [UMTS]/ [CDMA] din România".

The output file contains the following entities: Vodafone, România, Vodafone România, GSM, UMTS, CDMA. If an entity appears more than once, it will be found only once in the output file.

As an exemplification, here is a part of the XML output-file:

```
<entity type="company">Vodafone România</entity>
<category>conectivitate pentru serviciile de date</category>
<value ="1">bună</value>
```

- *recognizing categories, values and relationships with entities.* Considering the resulting files, once the previous phases have been completed, it will automatically extract the categories, values and relationships with entities using a set of rules (*regular expression*). These regular expressions use parentheses (round, square brackets) that form rules for constructing words. The most frequent use of regular expressions consists in recognizing if a string contains or not words or sub-string, that can be formed by that regular expression.

For instance: the string `p[oa]t` can be interpreted as *pot* and *pat*.

[13] http://gate.ac.uk/
[14] http://services.gate.ac.uk/annie/

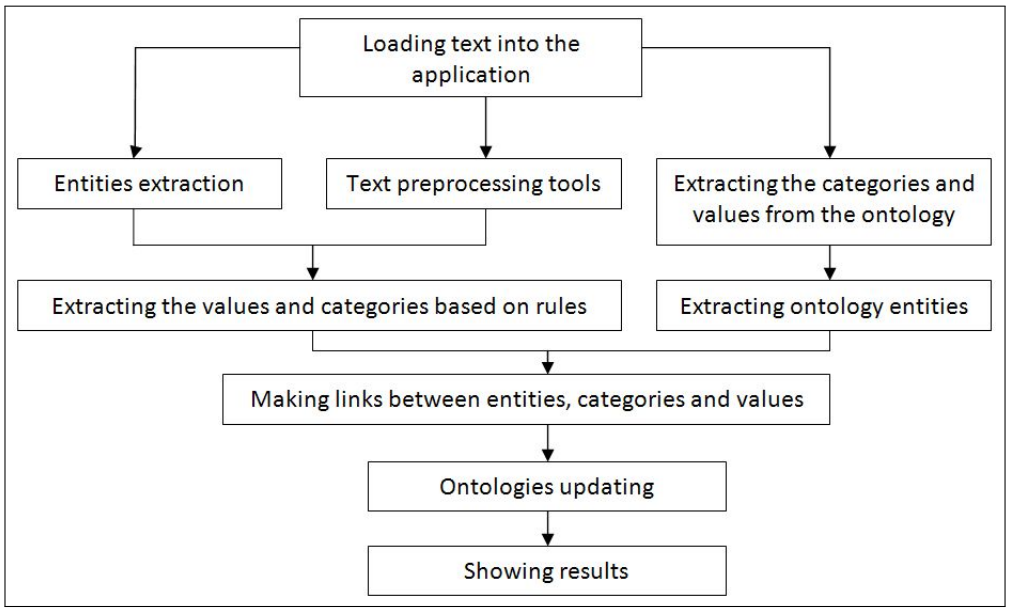

Figure 2: The architecture of the computational software

Basically, the tool completes the following steps:
- it identifies opinion words and phrases;
- it assigns to every positive or negative word a value (+1) for the positive one and (-1) for the negative one;
- the words which depend on context get also a value (0).

For instance: [Dacia] [Logan] este mai fiabilă decât orice [Opel].

```
<entity type="brand">Dacia Logan</entity>
<category>capacitatea sistemelor tehnice de a funcţiona </category>
<value ="1">fiabilă</value>
```

5 Conclusions and future work

This paper presents an automatic method able to detect and explain opinions on certain entities (peoples, companies, products, etc.) identified in a text, regardless of its nature (advertising, political, journalistic, etc.) based on a lexicon of opinions resulted from manual annotation (presented in other papers) of an initial corpus (consisting of opinion words and syntagmas). Moreover, in addition to this lexicon, we focused on the semantic role of negations and pragmatic connectors like "dar" ("*but*"). This application seeks to support the development of a complex lexical resource, necessary to interpret qualitative assessments found in any text. We are convinced that this analyze manner may be an important support for marketing managers, PR firms, politicians, online buyers, but, also, for specialists in NLP, linguistics, etc. Until now, we observed the fact that when a variable of neutralizing sentiments appears, it is not enough to cover only the summarizing operation of values for each opinion sentence. Because of that, we propose to add degrees of intensity and power in expressing opinions. In Romanian language, the superlative amplify semantically the convictions of the person who opines on an issue.

In the sentence - *Vodafone România oferă cea mai bună conectivitate pentru serviciile de date dintre toate reţelele mobile GSM/ UMTS/ CDMA din România.* - the word *bună* gets +1. The

superlative *cea mai* expands the scale of values. It can get the degree of positivity (or negativity). It depends on which word follows. So, *cea mai bună* gets (+2).

Also, due to pragmatic connectors, we have to give up on summarizing values.

Acknowledgments

In order to perform this research the first author received financial support from the Erasmus Mundus Action 2 EMERGE Project (2011 2576 / 001 001 - EMA2). I am also grateful to the NLP-Group@UAIC-FII for offering me support in using some tools for automatic interpretation of Romanian language.

Bibliography

[1] Ardeleanu, I. (2013); Extragerea de opinii din texte, lucrare de licenta coord. de prof.univ.dr. Dan Cristea, Universitatea Alexandru Ioan Cuza din Iasi.

[2] Dave, K.; Lawrence, S. and Pennock D.M. (2003); *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews* in Proceedings of WWW.

[3] Das, S.; Chen, M. (2001); Yahoo! For Amazon: *Extracting market sentiment from stock message boards* in Proceedings of the Asia Acific Finance Association Annual Conference (APFA).

[4] Ding, X., Liu, B. and Zhang, L. (2009): Entity discovery and assignment for opinion mining applications. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[5] Dragut, E.C., Yu, C., Sistla, P. and Meng, W. (2010): Construction of a sentimental word dictionary. In Proceedings of ACM International Conference on Information and Knowledge Management.

[6] Esuli, A.; Sebastiani, F. (2006); *Determining term subjectivity and term orientation for opinion mining* in Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, IT. Forthcoming.

[7] Gifu, D. (2012); Political Text Categorization in *Humanities and Social Sciences Review*, Vol. 1, No. 3, University Publications.net, USA, part of the paper presented in The International Journal of Arts and Sciences' (IJAS) International Conference for Academic Disciplines, Harvard University, Cambridge, Massachusetts, 27-31 May 2012.

[8] Gifu, D. (2013); *Temeliile Turnului Babel. O perspectiva integratoare asupra discursului politic*, Ed. Academiei Romane, Bucuresti.

[9] Habermas, J. (1962); Strukturwandel der Ofentlichkeit: Untersuchungen zu einer Kategorie der burgerlichen Gesellschaft. Neuwied, Luchterhand. [Trad. rom.: *Sfera publica si transformarea ei structurala*, Bucuresti, CEU, 1989.]

[10] Hatzivassiloglou, V. and McKeown K. R. (1997): Predicting the semantic orientation of adjectives. Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Madrid, ES, Association for Computational Linguistics.

[11] Ignat, E. (2011); RARE-UAIC (*Robust Anaphora Resolution Engine*), resursa gratuita pe META-SHARE, Universitatea "Alexandru Ioan Cuza" din Iasi, 2011.

[12] Jindal, N. and Liu, B. (2006): Identifying comparative sentences in text documents. In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval.

[13] Kamps, J., Maarten, M., R. ort.Mokken and Maarten de Rijke. (2004): Using WordNet to measure semantic orientation of adjectives in Proceedings of LREC-04, 4th International Conference of Language Resources and Evaluation, vol. IV.

[14] Liu, B. (2010); *Sentiment analysis and subjectivity.* Handbook of Natural Language Processing. N. Indurkhya and F.J. Damerau, eds.

[15] Liu, B. (2012); *Sentiment analysis and opinion mining.* Synthesis Lectures on Human Language Technologies, Morgan Claypool Publishers.

[16] Mihalcea, R.; Banea C.; Wiebe, J. (2007); *Learning Multilingual Subjective Language via Cross-Lingual Projections* in 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007).

[17] Pang, B.; Lee, L. (2008); Opinion mining and sentiment analysis in *Foundations and Trends in Information Retrieval*, 2.

[18] Pang, B.; Lee, L.; Vaithyanathan, S. (2002); *Thumbs up? Sentiment Classification using machine learning techniques* in Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing (Philadelphia, PA). Association for Computational Linguistics, Morristown, NJ.

[19] Pang, B.; Lee, L. (2004); *A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on minimum cuts* in Proceedings of the Association for Computational Linguistics.

[20] Simionescu, R. (2011); *POS-tagger hibrid*, lucrare de disertatie coord. de prof.univ.dr. Dan Cristea, Universitatea "Alexandru Ioan Cuza" din Iasi.

[21] Turney, P. (2002); *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification* of reviews in Proceedings of the Association for Computational Linguistics.

[22] Tong, R.M. (2001); *An operational system for detecting and tracking opinions in on-line discussion* in Workshop note, SIGIR 2001 Workshop on Operational Text Classification.