2011 International Nuclear Atlantic Conference - INAC 2011
Belo Horizonte,MG, Brazil, October 24-28, 2011
ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN
ISBN: **978-85-99141-04-5**

SENTIMENT ANALYSIS DOMAIN BASED ENVIRONMENT

Leonardo F. Koblitz[1] and Nelson F. F. Ebecken[2]

[1]Instituto de Engenharia Nuclear (IEN/CNEN)
Rua Hélio de Almeida nº 75 - CEP 21941-906
Cidade Universitária - Ilha do Fundão- RJ – Brasil
lfalcao@ien.gov.br

[2]COPPE/Universidade Federal do Rio de Janeiro
Avenida Horácio Macedo nº 2030
Prédio do CT, Bloco B - Sala 100 - CEP 21941-914
Cidade Universitária - Ilha do Fundão - RJ – Brasil
nelson@ntt.ufrj.br

ABSTRACT

In this article, the Sentiment Analysis, which main task is to understand the polarity of the opinions expressed in the documents, is used to obtain a better understanding of how people express themselves about a particular subject or domain.
Such understanding is important because it may provide input in the formulation of policies and actions to be taken in relation to a product or service or to the perceptions related to issues or people.
As more people put their thoughts and opinions on a variety of services available on the Web microblogging sites like Twitter, social networks and forums have become a common way for them to express themselves.
There, they put in a spontaneous, free and in real time over different views on the issues.
However, each domain or Internet service has its own peculiarities. Some jargon is specific to a domain. Certain slang or characteristics of services for people putting their opinions differ significantly, which undermines the use of machine learning systems previously developed for other domains.
To validate the proposed methodology a corpus of nuclear texts was assembled, annotated and analyzed. After that, a system was developed to process these data.
The results thus obtained provide important information on the identification of key stakeholders and the views expressed on what subjects. With this better understanding of the fears, prejudices and expectations of people regarding the nuclear area, new strategies can be traced to improve the perception of them as the nuclear area.

1. INTRODUCTION

We present JULGAR an environment to perform sentiment analysis of texts as a function of the domains under analysis.
The importance of this strategy lies in the fact that the Internet is becoming a common way for people to express their opinions and sentiments freely and spontaneously and also because of the appearance at all times of new services with different characteristics in which people are encouraged to give their opinion on numerous domains.
To meet these demands, various softwares and tools are being developed to analyze either the sentiments or opinions expressed in the texts. Each of these systems however, has its own and distinct characteristics.

ISBN: **978-85-99141-04-5**

In order to carry out this work a corpus with texts referring to the nuclear area was built and annotated as well as a corpus with messages posted on the microblogging service "Twitter"[1]. The following in the next section are some related products used to perform sentiment analysis.

2. RELATED WORK

Bartlett and Albright [1] analyzed several strategies to perform sentiment analysis using the SAS software TextMiner, which is a major supplier in this action.
However, they realized that the strategies tested did not produce the expected results, because the systems evaluated are closely related to the domain they were designed to meet.
Pak and Paroubek [2] have developed software to perform sentiment analysis of the messages contained in "Twitter", called "tweets." They have used as an indicator of the polarity of the sentiments of the messages posted in the construction of the corpus, the type of "emoticons"[2] contained in the messages.

The company SentiMetrix offers software which has a module to analyzing sentiments or reputations. Currently this system is being used as part of a government project that aims to help families whose members have returned from conflicts in Iraq and Afghanistan and have emotional depressive disorders.
However, the examples given above are not completely reliable and can therefore be inaccurate. The main reason for this inaccuracy is that the systems reviewed were developed specifically to meet a given domain and thus have its field of application restricted.

Moreover, these systems don't have a strategy integrated to them to incorporate the specific characteristics of a particular domain and thus increase their range.

With this in mind, this article aims to define, implement and validate a strategy to perform sentiment analysis based on domain.

As a starting point for this work the nuclear area was chosen. The choice of this domain was due to the fact that this area presents a high degree of controversy. This choice helped to obtain the necessary understanding to develop a sentiment analysis domain based environment.
As support to validate this strategy a computational environment called JULGAR was developed. This environment can be configurable depending on the application or domain being evaluated.

These are some of the ways JULGAR can be configurable: for instance, the incorporation of polarity and subjectivity classifiers developed and tested outside this environment, the definition of various types of lists, such as "stop words" list, lists of words or phrases with positive or negative polarity for the domain being analyzed.

---

[1] http://twitter.com
[2] An emoticon is a facial expression pictorially represented by punctuation and letters, usually to express a writer's mood.

ISBN: **978-85-99141-04-5**

The environment JULGAR also allows multiple types of content analysis of the texts being analyzed.

This strategy covers the following: the corpus annotation process, the steps necessary for creating notes according to the area, creation of semantic lexical and the development and validation of the classifiers.

3. DOMAIN ADAPTATION STRATEGY

A scarcity of labeled data is faced when migrating to a new domain. Achieving these new data is expensive and time consuming. The traditional algorithms of machine learning require both labeled and unlabeled data extracted from the same distribution.

Adaptation domain refers to the process of adapting a trained extraction model developed for a specific domain to another related domain with unlabeled data only.
That is, the adaptation domain is different from the semi-supervised learning, which assumes that the labeled and unlabeled data are drawn from the same domain. It is assumed that although the domains are different, the conditional distribution of classes in the two domains remains unchanged, i.e., $P(y \mid x) = P(y_i \mid x_i)$ [3].

Various methods of domain adaptation [3, 4, 5] have been proposed to overcome this lack of labeled data. The basic idea of domain adaptation methods is to choose examples of representation of a known domain which enables them to get closer to the distribution of the new domain.

These are some of the domain adaptation techniques:

- Select the relevant examples;
- Remove irrelevant features;
- Add related features;
- Using found regularities (properties) in the domain records that were adapted to fit the classifiers.

One can summarize the strategy proposed in this work through the following steps:

- Creation of a annotated corpus;
- Develop routines for extracting and processing the domain-specific annotations;
- Creation of lexical semantics, domain-specific expressions and lists with expressions of subjectivity and lists with polarity of words;
- Construction of the classifier
    - Choice of features;
    - Choice of algorithm to be used;
    - Parameterization of the configuration file of the classifier;
    - Evaluation of results;
    - Save the classifier for later use.
- Specialization for the domain to be analyzed

ISBN: **978-85-99141-04-5**

4. SYSTEM OVERVIEW

The JULGAR system was developed using the JAVA programming language running on Eclipse platform. The entire process of document annotation, as well as the development of the classifiers was conducted by using the GATE software and then transferred to the system JULGAR.

The GATE software project is the largest open source for natural language processing. The software GATE offers software developers a complete infrastructure for developing software components for different types of applications, such as: creation of linguistic features, creation of ontologies and building classifiers.

JAPE language (software embedded in the GATE) was used to extract, create and process annotations in the texts analyzed. JAPE language allows the creation of regular expressions to be used on the annotations contained in the documents. This language uses LHS / RHS rules, where LHS is the pattern to be found and RHS is the action to be performed.

After annotations were created subjectivity and polarity classifiers were developed. This was done through the component Batch Learning PR where the algorithm is defined to be used and what linguistic features are considered for processing. Figure 1 shows the development flow adopted in this work.

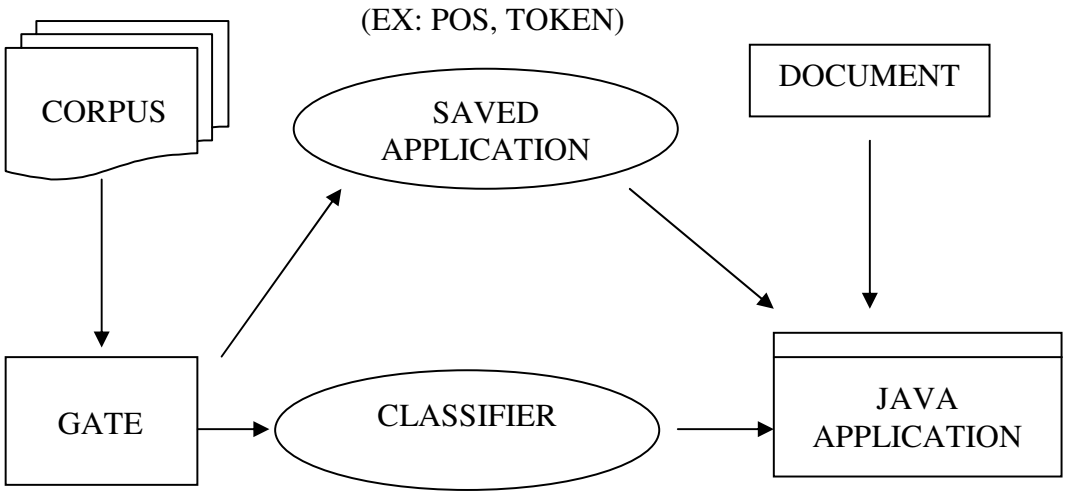

Figure 1. Development flow of the JULGAR system.

***ISBN:* 978-85-99141-04-5**

5. CORPUS ANALYSIS

5.1 The nuclear area

Although the nuclear area is so present in our lives, as for example, power generation, medicine and many other fields, it suffers a stigma around the world, that relates itself directly to its use in nuclear weapons or radioactive accidents occurred in the past. Several studies have been done to better understand how people think, what motivates them and what actions could be done to mitigate these negative perceptions of the nuclear area.
As a conclusion we can realize that the perceived risk in the use of nuclear energy depends mainly on personal values, experiences and their social groups [6, 7, 8, 9].

These are the main conclusions:

- Young people are the most promising group to address efforts to improve the perception of the nuclear area;
- It is important to avoid excessive repetition of some words like “safety”, “security” or “peaceful uses” in presentations[3];
- The more you teach people about nuclear energy the more they can be against it;
- People are influenced by political views and gender.

However, as studies suggest, people's perceptions can be improved among other actions, by segmenting the audience according to their dominant mental models.
Understanding how feelings are expressed and the identification of external factors that influence them are very important to understand, predict and propose actions to mitigate the rejection of people to new technologies.
This learning can therefore be used not only in the nuclear area, but also to serve as a guide when there is the need to propose actions referring to the use of new technology.

5.2 Corpus annotation

To understand the nuclear area three corpora were collected. The first corpus was composed with texts for nuclear area and the second against. But the last one had no opinion at all. The texts that make up for the three corpora were taken from various websites, blogs and Internet forums.

The annotation of corpora was manually done using the scheme proposed by Wilson [10] and it took two and a half months to be accomplished. Two hundred and eighty one documents were noted (for and against). The annotation process performed identified several expressions and assigned values for the characteristics of these expressions.
According to Wiebe [11] the goal of the annotation scheme is to represent the mental and emotional states contained in internal documents and distinguish factual information from subjective information.

[3] The use of these words make people wonder if nuclear area is so safe, secure and so on.

ISBN: 978-85-99141-04-5

Other annotations were also created based on the sociolinguistic observations listed by Spertus [12] who identified in his work a series of phrases that convey a negative idea even though the words in these phrases do not contain negative connotations. For example, sentences with "your so called" or "your so-called" like "Read through your so called evidence" has 50% of chance to be a flame.
The use of these expressions helped build better classifiers to identify expressions of subjectivity.

6. CREATION OF A SEMANTIC LEXICON

In order to perform automatic annotation of words with positive and negative polarity a subset of the list of words that express subjectivity created by Wilson and Wiebe [13] was used as a starting point.
Words with neutral polarity and indicating low subjectivity have been deleted to increase the accuracy of the classifier of subjectivity. The words classified with high subjectivity have a great indicator of subjectivity, as opposed to those classified as low.
Two lists were therefore created with positive and negative polarity words which were used in the annotation process.

Afterwards the software AUTOSLOG-TS [14] and BASILISK [15] were then configured for the creation of semantic lexicon for the nuclear domain.
The AUTOSLOG-TS has as input two sets of texts. The first one is relevant to the analyzed domain but the other one is not. A very large set of extraction patterns is produced as result and should be checked manually at the end.
In order to find the new values of each semantic category the extraction patterns produced in the last step are associated to an initial list of semantic categories as input of the software BASILISK.

A list of words that typically appear in texts against nuclear area was also created. As an example we have Chernobyl, the Russian nuclear power plant where a nuclear accident occurred on 04/26/1986. The lists from previous steps were put together to produce the semantic lexicon used in the annotation of texts. Figure 2 shows the development flow of a semantic lexicon for a specific domain.

7. CLASSIFIERS CREATION

The strategy suggested by Pang and Lee [16] was adopted to find out the polarity of a sentiment being expressed in a text, which performs this task through two steps. The first step classifies the sentences in subjective and objective. The following step, the objective sentences, that have neutral polarity, are discarded and then only the subjective sentences are processed to obtain the polarity of the text.
Then two classifiers were developed. The first developed classifier classifies the sentences in subjective or objective.

ISBN: **978-85-99141-04-5**

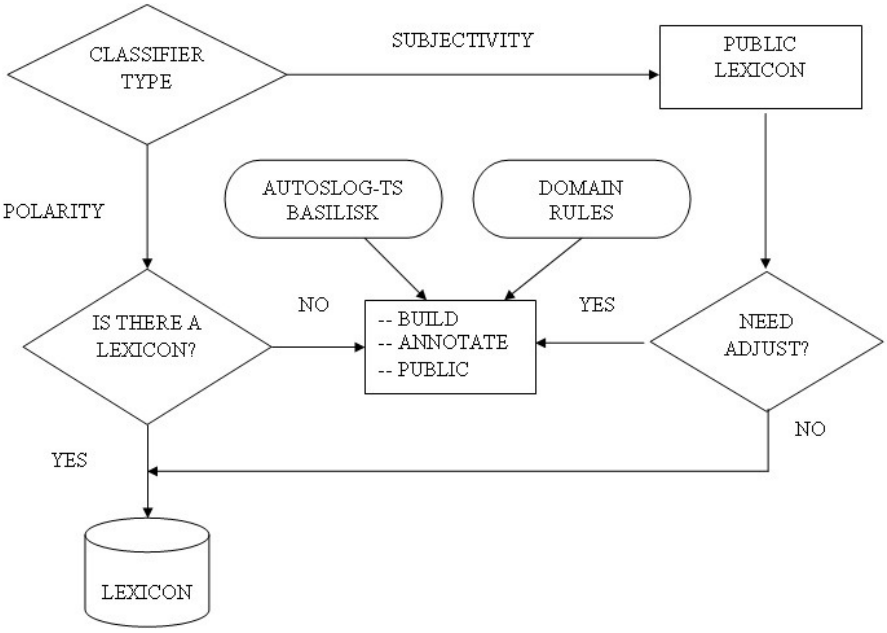

Figure 2. Development flow of a semantic lexicon for a specific domain.

The second polarity classifier is applied only on subjective sentences. This strategy allows achieving better results than applying directly a single polarity classifier on texts. To improve the accuracy of the classifiers various configurations of linguistic features, semantic lists (called "gazetteers" in the GATE) and "n-grams" were tested. The best results for both classifiers were obtained with SVM algorithms and with the use of "unigrams."

8. RESULTS

A second experiment using the corpus produced by Pak and Paroubek [2] was used in this article to validate the strategy. This corpus is composed by 300,000 texts taken from "Twitter", forming three corpora evenly divided as a result.
The first one expressing positive sentiments, the second one negative sentiments and the last one with only objective texts.
The subjectivity and polarity classifiers evaluation was performed within the GATE environment which provides various metrics on the training conducted, for example, "recall" and "precision" among others.

The classifiers that were built into the system JULGAR have been tested with some previously selected set of documents and also with recent texts taken from sites such as the anti-nuclear organization Greenpeace. Table 1 presents the results of the subjectivity classifiers on the nuclear corpus.

ISBN: 978-85-99141-04-5

Table 1. Results of the subjectivity classifiers on the nuclear corpus

algorithm	n-grama	Features	Precision	Recall	F1
SVM	1	sent_size	0.7061	0.7061	0.7061
PAUM	1	sent_size	0.5053	0.5053	0.5048
SVM	2	sent_size	0.7911	0.6973	0.6973
SVM	1	spertus	0.7252	0.7252	0.7252
PAUM	1	spertus	0.5012	0.5009	0.5011

* Where "sent_size" is the length of the sentence and "spertus" is an annotation based in the sociolinguistic observation proposed by Spertus [12]. Both of them are used together with others linguistic features.

In addition to the calculation of the polarity of the sentences the system JULGAR makes various types of content analysis. As an example this system identifies co-occurrence of words in the corpus, and also co-occurrence of words in relation to a specific list.

Table 2 shows the words that most co-occur with names of countries. This analysis showed that the word "Clamshell" which is an anti-nuclear organization appears frequently related to France and then later can be used as an indicator of negative polarity.

Table 2. List of words that co-occur with the names of countries

France	China	USA
nuclear=92	nuclear=65	national=7
energy=33	energy=25	victory=5
reactors=32	reactor=23	energy=5
reactor=32	reactors=20	atomic=4
French=31	French=15	reactor=3
Clamshell=23	plants=13	likely=3
radioactive=18	electricity=12	guarantee=3
industry=18	construction=10	campaign=3

The table 3 presents two types of private states[4] learned from the manual annotation process on the nuclear corpus.

[4] It is a state that is not open to objective observation or verification like an opinion.

***ISBN:* 978-85-99141-04-5**

Table 3. Examples of private states extracted from the nuclear corpus

Types of private states	
Expressive-subjectivity	Direct-subjective
absolutely urgent	can afford
brought back	censored
sustainable culture	collapse
delays and regulatory battles persist	gearing up
History sounds a cautionary note	opposed
more clean	promoted
no tomorrow	unacceptable

9. CONCLUSIONS

This work had as a main contribution to establish an environment to perform sentiment analysis on a new domain. The need to promptly meet the new demands arise all the time with new services being made available on the Internet that require an environment that allows easy configuration to reduce the time to develop an application for this purpose. Moreover, the system JULGAR provides content analysis of texts. This type of information can serve as feedback to aid the construction of semantic lexicons and, consequently, the improvement of the classifiers to be used in other domains.

The strategy proposed in this paper was successfully tested when this system was rapidly configured to process another domain (a set of messages taken from "Twitter"). To do this it was only necessary to train the classifiers and to incorporate the external information of the new domain being analyzed.

In addition to contributions already mentioned, another important one was the adaptation and inclusion of the rules proposed by Spertus [11] to process both the nuclear and the "Twitter" domains. The use of these rules provided a better understanding of the texts and therefore a better performance of the classifiers employed.

The computer programs AUTOSLOG-TS and BASILISK should be integrated with the JULGAR system to improve the analyze of a new domain. In this case, the new entries learned would be used as entries for the polarity lexicons.
As a future, it work would be important to migrate the JULGAR system and linguistic processing modules for the Portuguese language.

ACKNOWLEDGMENTS

Special thanks to Dr. Antonio Cesar Ferreira Guimarães and to the various users and developers of GATE for their helpful discussions.

ISBN: **978-85-99141-04-5**

REFERENCES

1. J. Bartlett and R. Albright, "Coming to a Theater Near You! Sentiment Classification Techniques Using SAS® Text Miner", SAS Global Forum 2008, SAS Institute Inc., Cary, NC, USA (2008).
2. A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.", Proceedings of the Seventh Conference on International Language Resources and Evaluation, 1320-1326 (2010).
3. R. Gupta and S. Sarawagi, "Domain adaptation of Information Extraction Models", ACM SIGMOD, Volume 37, Issue 4, 35-40., Canadá (2008).
4. A. Aue, and M. Gamon, "Customizing Sentiment Classifiers to New Domains: a Case Study". Proceedings of International Conference on Recent Advances in Natural language Processing (RANLP), [7 p.], Borovets, BG, (2005).
5. T. McIntosh, and J. Curran, "Weighted Mutual Exclusion Bootstrapping for Domain Independent Lexicon and Template Acquisition", Proceedings of the Australasian Language Technology Workshop, [p.9], Hobart, Australia, (2006).
6. J. Ribeiro, A. Barroso and K. Imakuma, "The Communication of the Value and Public Acceptance of Nuclear Plants", Proceedings of ICAPP 2007, 13-18, France (2007).
7. J. Costa-Font, C. Rudisill and E. Mossialos, "Attitudes as Expression of Knowledge and "Political Anchoring": The Case of Nuclear Power in the United Kingdom", Risk Analysis, vol. 28, n.5, 1273 – 1287 (2008).
8. J. Junior, A. Barroso et al., 2007, "News and its influence on the viability of nuclear power plants deployment – a modified epidemiological model for news generation", Proceedings 2007 International Nuclear Atlantic Conference – INAC 2007, [p.7] Brazil (2007).
9. A. Kugo et al, "Study on risk communication by using Web system for the social consensus toward HLW final disposal", Progress in Nuclear Energy, vol. 50, 700 – 708 Japan (2008).
10. T. Wilson, "Fine-Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States", D.Sc. Dissertation, University of Pittsburgh, Pittsburgh/USA (2007).
11. J. Wiebe, "Instructions for Annotating Opinions in Newspaper Articles", Technical Report TR-02-101, University of Pittsburgh, Pittsburgh, PA. (2002).
12. E. Spertus, "Smokey: Automatic Recognition of Hostile Messages", Proceedings of Innovative Applications of artificial Intelligence (IAAI), 1058-1065, Providence, Rhode Island (1997).
13. T. Wilson, and J. Wiebe, "Annotating Attributions and Private States", Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, 53-60, NJ, USA, (2005).
14. E. Riloff and W. Phillips, "An Introduction to the Sundance and AutoSlog System", Technical Report UUCS-04-015, School of Computing, University of Utah, USA (2004).
15. M. Thelen and E. Riloff, "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts", Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, USA (2002).
16. B. Pang and L. Lee, "A Sentimental Education: Sentiment analysis using subjectivity summarization based on minimum cuts", Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, 271-278, Barcelona, ES (2004).