IA-Regional-Radio - social network for radio recommendation.

Grzegorz Dziczkowski[1], Lamine Bougueroua[2], and Katarzyna Wegrzyn-Wolska[2]

[1] Ecole des Mines de Paris, 77305 Fontainebleau, France,
[2] Ecole Superieur d'Ingenieurs en Informatique et Genie des Telecommunicatiom, 77-215 Avon-Fontainebleau Cedex, France,
(grzegorz.dziczkowski lamine.bougueroua katarzyna.wolska)@esigetel.fr

Summary. This chapter describes the functions of a system proposed for the music hit recommendation from social network data base. This system carries out the automatic collection, evaluation and rating of music reviewers, the possibility for listeners to rate musical hits and recommendations deduced from auditor's profiles in the form of regional internet radio. First, the system searches and retrieves probable music reviews from the Internet. Subsequently, the system carries out an evaluation and rating of those reviews. From this list of music hits the system directly allows notation from our application. Finally the system automatically create the record list diffused each day depended form the region, the year season, day hours and age of listeners. Our system uses linguistics and statistic methods for classifying music opinions and data mining techniques for recommendation part needed for recorded list creation. The principal task is the creation of popular intelligent radio adaptive on auditor's age and region - IA-Regional-Radio.

1 Introduction and issue

Social networking sites are a global phenomenon. For the hundreds of millions of people worldwide who belong to sites like MySpace, Facebook and YouTube, social interaction in cyberspace has become an indispensable part of their lives.

Nowadays the internet is an essential tool for the exchange of information on a personal and professional level. The web offers us a world of prodigious information and has evolved from simple sets of static information to services which are more and more complex. With the growth of the Web, internet radio, recommendation tools and e-commerce has become popular. Many websites offer on line services or sales and propose object ratings to their users, for music, films, and products for example. With globalization, the product choice is too much diversified, therefore, users are not aware of the availability of products. For this reason, prediction engines were developed to offer the

user alternative products. Generally, the influences of others is important for opinion making.

Prediction engines' algorithms are based on the experience and opinion of other users. In order to do those engines, we need to have an extremely large user profile base. In our case - internet radio, products furnish by our system are musical hits. In this case, auditors are not supposed to make the validation process of listened musical hit. The need of such system is justified by increasing value of radio auditors. For example, in France there are 42 millions of auditors each day, it represent 8 people out of 10 older then 13 years old [19]. The value of internet radio listener drew over 1.2 additional millions of listeners aged 13 and over last year [Mediametrie 126 000 Radio 2007-2008] [19].

The general objective of our system is to furnish the intelligent internet radio which needs to be programmed first but which interacts with the taste of listeners from the same region and of the same age. The vote of the listeners will directly affect the single rotation frequency. Another advantage of our system is that the prediction are not made for each auditor but for the group of auditors from same region and same age. Like this, we present also musical hits which have not been discovered yet for singular listener. In addition, our method is more useful for new auditors which didn't evaluate too many musical hits. The time at which auditors connect to the Internet to listen to the music is arbitrary. It is possible that a musical hit is very appreciated in the morning but not at night. Therefore, we look at voting time to understand better the tastes of users.

Most of radios are using automation radio software. In fact, it gives the possibility to make playlists, to play all type of song (jungle, advertisement, music, interviews, live, ...) and that with just a simple computer. In France for example, we have a radio totally automated, "Chante France". There are no speakers, only a computer which play songs. Some products exist in radio automation software. These products are used by associative or personal radios. There is also a little list of professional products used by national radios.

The player Zarasoft has no database, there is no possibility to program and regroup onto categories. But there is the possibility to program events at specific time. It detects silence at the end of track. It is every time necessary to add manually the playlist, or to select a directory, and zararadio selects randomly the songs to play, but there is no possibility to show the time remaining for introduction. We can describe this software as semi-automatically.

Zradio is a freeware, developed by a local radio RMZ at Poitier before 2003. It has the possibility to play Jingles, advertisements and to be program in a format by using a database. Winamp is a simple free media player developed since 1997. Contrary to Zararadio, it can not select randomly a song by directory. There is no automation because we need to manually program all the song we want to hear. Radugo exists since before October 2001. We need to manually add all songs. There is a possibility to create list think, log files automatically generate, passwords protection, scheduled events and

silence detection.

The player DRS2006 includes automatics playlist editor. With the plug-in broadcast, it's possible to directly broadcast the stream on the Internet. It includes a database, a playlist editor, a studio Radio and studio DJ, and some tolls about database. Easyradio exists since 1998. It's the most complete software in radio automation for little radios. It's simple of use, it has the possibility to define some time formats, and the playlist are generated automatically. This software is very complete, and is recommended for DJs, radios, expositions. The editor of Sam Broadcaster has been created in 1999. It seems to be complete software. It integrates a web broadcaster. We can make some graphs with the statistical number of listeners. It gives an html output that allows the update of the song play, for example, for a website.

All softwares have its own advantages and disadvantages. But all the advanced solutions do not include an automatic update of the song frequency based on the vote of the listeners.
System presented in this chapter searches and retrieves probable music reviews from the Internet. Subsequently, the system carries out an evaluation and rating of those reviews. From this list of musical hits the system directly allows notation from our application. Finally the system automatically create the record list diffused each day depended form the region, the year season, day hours and age of listeners. Our system uses linguistics and statistic methods for classifying music opinions and data mining techniques for recommendation part needed for recorded list creation. The principal task is the creation of popular intelligent radio adaptive on auditor's age and region - IA-Regional-Radio.

2 General system architecture

In this section, we make a description of the architecture. Our objective is to make internet radio which interacts with the taste of listeners. Principal modules of our architecture are [Figure 1]: research and collect of reviewers on Internet, attribution of a mark for each review and storage of interesting information in database. We also store mark collected from internet radio website. The most important part of our modular architecture system is *Opinion marking module* described in section 6. The recommendation system is based on information from *opinion marking module* to understand better tastes of users.

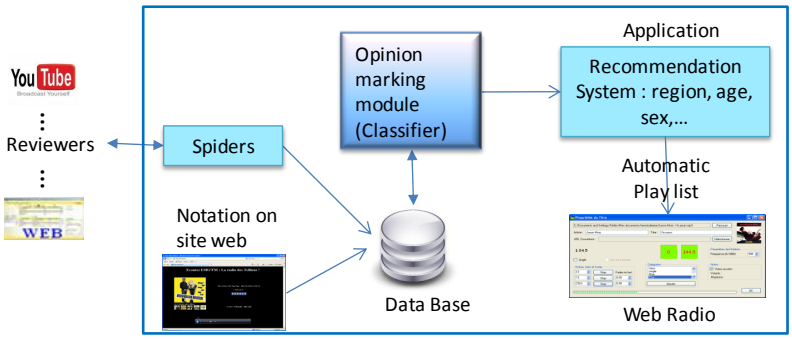

Fig. 1. Application architecture.

The last module is used to generate lists of songs depending on several criteria : region and age of listener, recommendation results, hours of broadcasting.

The first part will be to make a database and an interface to add songs on it. The field in database will be essentially: title of song, artist name, category, frequency, listener age, listener region and some others information [Figure 2].

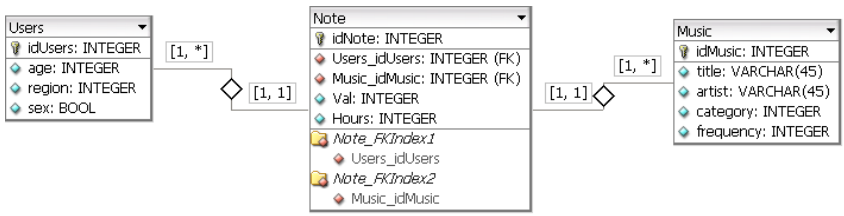

Fig. 2. Data base.

The next part is to collect information about musical hits on internet (youtube, ...). In fact, we use spider which is the automated and methodological traverse and index of web pages for subsequent search purposes. The application has a web site to receive request of web listeners, which result in a note or comment on musical hits.

All relevant information are, then, recovered either directly from the application, or by the spiders. Finally, informations are stored in a database. After collecting the reviews, we will assign notes by using the classifiers. The classifiers provide ratings from users' feelings. The classifier uses three different methods for assigning a mark to the reviews. Those methods are based on different approaches of corpus classification. The notes of the classification will be stored in a database.

The recommendation system, relying on the notes in the database and region, sex and age of the user, determine the most appropriate musical hits

to broadcast. Each day, the application auto-generate playlists by using the playlist scheme generator model [Figure 3]. It consists to select the sound to play by respecting the restrictions and the playlist scheme according to the song frequency in the database.

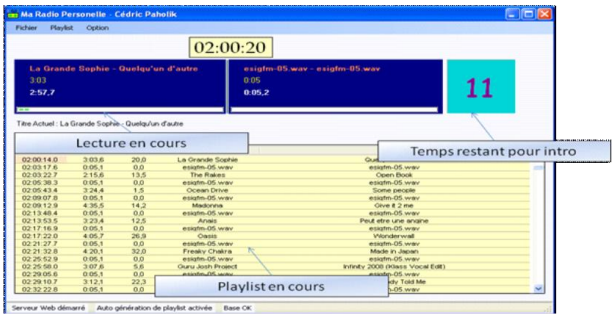

Fig. 3. Web radio Application - server.

The web radio is composed of two parts. The first part is the studio player interface. It represents the server side. It displays the list during playback and the remaining time for start playing musical hits.

The second part is the web interface. It represents the listener interface. It allows listener to view the jacket and title name being circulated, the last five titles available, the vote for the title currently broadcast and the possibility to obtain additional information on a title.

3 Text Mining

The text categorization *TC* is now applicable in many different contexts. Document indexation is based on a lexicon, a document filtration, an automatic generation of metadata, a suppression of words ambiguity, the settlement of hierarchical catalogues of Web resources, and, in general, every application, which needs document organization or selective processing and document adaptation [16], [23].

The Machine Learning *ML* describes a general inductive process, which automatically constructs a text classifier via the learning, from a series of preclassified documents or from characteristics of interest categories. Text Mining *TM* is a set of informatics processing, which consists of extracting knowledge in terms of innovative criteria or in terms of similarities in texts produced by human beings for human beings. A field using *TC*, *ML* or *TM* techniques is, in particular, the field of sentimental analysis [4], known as Opinion Mining [5]. The research in this field covers different subjects, in particular the learning of words' or expressions' semantic orientation, the sentimental analysis

of documents and opinions and attitudes analysis regarding some subjects or products [18], [15].

3.1 Representation of documentary corpus

Texts in natural language can not be directly interpreted by a classifier or by classification algorithms. The first linguistic units representing the sense are words' lemma. The recognition of those linguistic units requires carrying out a linguistic preprocessing of the text's words. The number of words characterizing a document corpus can be really wide. Therefore, it is necessary to conserve a subgroup of those words. This filtering relies at the root on words occurrence frequencies in the corpus.

Other approaches are using not words but group of words, eventually sentences such as linguistic units, describing the sense. Thanks to this approach, we have an order relationship between words and words' co-occurrences. The inconvenient is that the frequency of group of words apparition can not offer reliable statisticals because the great number of combinations between words creates frequencies which are too wick to be exploited.

Another approach to represent the documentary corpus is the utilization of the n-grams technique [26]. Those methods are independent from the language, however neither the segmentation in linguistic units, nor pre-treatments as filtration and lemmatization are necessary.

If we are using words such as linguistic unit, we notice that different words have common sense or are simply another form of conjugation. Therefore, a processing named stemming has to be carried out. It is a processing, which proceed at a morphologic analysis of the text [24]. The processing, which needs a more complex analysis than stemming, is lemmatization, which is based on a lexicon. A lexicon is a set of lemmas with which we can refer to the dictionary. Lemmatization needs to carry out in addition a syntax analysis in order to resolved ambiguities. Therefore, it conducts a morphosyntactic analysis [25].

The role of textual representation is represented mathematically in a way that we can carry out the analytic processing, meanwhile, conserving at a maximum the semantic one. The indexation process itself consists in conducting a simple complete inventory of all corpus lemmas. The next step is the selection process of the lemma, which will constitute linguistic units of the field or vector space dimension of the representation of documentary corpus.

3.2 Classification techniques

The classification procedure is automatically generated from a set of examples. An example consists in a description of a case with the corresponding classification. We dispose, for example, a database of patients' symptoms with the status of their respective health state, as well as the medical diagnostic of their sickness. The training system must then, from this set of examples, extract a classification procedure, which will, with a view of patients' symptoms,

establish a medical diagnostic. It is a matter of inducing a general classification procedure taken from examples. The problem is therefore an inductive problem. It is a matter of extracting a general rule from observed data.

Bayes' Classifier

The probabilistic classifier interpret the function $CSV_i(d_j)$ in terms of $P(c_i|\mathbf{d_j})$, which represents the probability that a document is represented by a vector $\mathbf{d_j} =< w_1, j, ..., w_{|T|j} >$ of terms, which belongs to ci, and determine this probability by using the Bayes' theorem, defined by:

$$P(c_i|\mathbf{d_j}) = \frac{P(c_i)P(\mathbf{d_j}|c_i)}{P(\mathbf{d_j})}. \quad (1)$$

Where $P(\mathbf{d_j})$ is the probability that a document choose at random, has the vector $\mathbf{d_j}$ for its representation; and $P(c_i)$ is the probability that a document choose at random, belongs to c_i. The probability estimation $P(c_i|\mathbf{d_j})$ is problematic, since the vector number $\mathbf{d_j}$ possible is too high. For this reason, it is common to make the hypothesis that all vector coordinates are statistically independent. Therefore:

$$P(\mathbf{d_j}|c_i) = \Pi_{k=1}^{|T|} P(w_{kj}|c_i)). \quad (2)$$

Probabilistic classifier, which are using this hypothesis are named Nave Bayes' classifier and find their usage in most of probabilistic approaches in the field of text categorization [29], [17].

Calculation of a classifier by the SVM method

Support Vector Machine methods has been introduce by Joachims [13], [14], [6], [35]. The geometrical SVM method can be considered as an attempt to find out between surfaces $\sigma_1, \sigma_2, ...$ of a dimension space $|T|$, what is separating examples of positive training from negative ones. The set of training is defined by a set of vectors associated to the belonging category: $(X_1, y_1), ..., (X_u, y_u)$, $X_j \in R^n, y_j \in \{+1, -1\}$ with:

- y_j represents the belonging category. In a problem with two categories; the first one correspond to a positive answer ($y_j = +1$) and the second one correspond to a negative answer (y_j = -1)
- X_j represents the vector of the text number j of the training set.

The SVM method distinguish vectors of positive category from those of negative category by a hyperplane defined by the following equation: $W \otimes X + b = 0, W \in R^n, b \in R$.

Generally, such a hyperplane is not unique. The SVM method determines the optimal hyperplane by maximizing the margin: the margin is the distance between vectors labeled positively and those labeled negatively.

Calculation of a classifier by the tree decision method

A text classifier based on the tree decision method is a tree of intern node, which are marked by terms, branches getting out of node are tests on terms, and the leaves are marked by categories [20]. This classifier classify document of the test d_j by testing recursively the weight of intern node of the vector d_j , until a leaf is reach. The knot's label is then attributed to d_j . Most of those classifiers use a binary document representation and therefore are created by binary trees.

A method to conduct the training of a decision tree for the category c_i consist in verifying if every training example have the same label (c_i or $\bar{c_i}$). If not, we will select a term t_k, and we will break down the training set in document categories, which have the same value for t_k. Finally, we create subtrees until each leaf of the tree generated this way contain training examples attributed at the same category c_i, which is then choose as the leaf label. The most important stage is the choice of the term of t_k to carry out the partition [20].

Neural Network

A text classifier based on neural network *NN* is a unit network, where entry units represent terms, exit units represent the category or interest categories, and the weight of the side relating units represents dependency relationships. In order to classify a document of test dj, its weights wkj are loaded in entry units; the activation of those units is propagating through the network, and the value of the exit unit determines the classification decision. A typical training way of neural network is retro propagation, which consist in retro propagating the error done by a neuron at its synapses and to related neurons. For neural networks, we are usually using retro propagation of the error gradient [3], which consist in correcting error according to the importance of the elements, which have participated to the realization of those errors.

4 Sentiments Analysis

4.1 The complexity of opinion marking

In order to determine the complexity of opinion marking, we are going to take an example of a review. The example is:

> "Yeah, Beautiful girl. I've only met 2 people in real life and 1 person on the net who hates this musical hit. My favorite song ever!"

As we have noticed, the review is composed of three phrases, which have opposite polarity. Even though, we can easily deduct that the first sentence is the movie title, *Beautiful girl*, we will have two subjective phrases but hard

to mark correctly. The last phrase is rather easy to mark: *"My favorite song ever!"*. However, there is a problem for the marking of the phrase: *I've only met... who hates this musical hit*, because a statistical study shows us that the polarity is negative for this phrase but in fact the polarity is positive and with high intensity.

Sentiments can often be express in a subtle manner, which creates a difficulty in the identification of the document units when considering them separately. If we consider a phrase, which indication a strong opinion, it is hard to associate this opinion with keywords or expressions in this phrase. In general, sentiments and subjectivity are highly sensitive to the context and dependent of the field.

Moreover, on Internet, everyone is using its own vocabulary, which adds difficulties to the task; even though it is in the same field. Furthermore, it is very hard to correctly allocate the weight of phrases in the review.

It is not yet possible to find out an ideal case of sentiment marking in a text written by different users because it does not follow a rule and it is impossible to schedule every possible case. Moreover, frequently the same phrase can be considered as positive for one person and negative for another one.

4.2 Detection of subjective phrases

For many applications, we have to decide if a document contains subjective or objective data and to identify which parts of the document are subjective in order to be able, then, to process only the subjective part.

Hatzivassiloglou and Wiebe [12] have demonstrated the phrases' orientation based on the adjectives' orientation. The objective was to establish if a given phrase is subjective or not by evaluating adjectives of this phrases [34], [1]. Wiebe et al. [31] present a complete study of the recognition of subjectivity by using different indications and characteristics (the results comparison by using adjectives, adverbs and verbs in taking in account the syntax' structure like for example the words' location).

Another approach made by Wilson et al. [33] has proposed an opinion classification according to their intensity (the opinion strength) and according to other subjective elements. When other researches have been made on the distinction between subjectivity and objectivity or on the difference between positive and negative phrases, Wilson et al. have classified the opinion and emotion strength expressed in individual clauses. The strength is known as neutral when it corresponds to the absence of opinion and subjectivity.

Recent works consider as well relationship between the ambiguity in the words sense and in the subjectivity [32]. The subjectivity detection can also be done thanks to classification techniques.

4.3 The opinion polarity and intensity

The classification of the opinion polarity consists in a document classification between positive and negative status. A value called semantic orientation has

been created in order to demonstrate words' polarity. It varies between two values: positive and negative and can have different intensity level. There are several calculation methods of the words semantic orientation. Generally, the semantic orientation method of the associations SO-A is calculated as a measure of positive words association less the measure of negative words association:

$$SO - A(word) = \Sigma_{p \in P} A(word, p) - \Sigma_{n \in N} A(word, n) \tag{3}$$

Where A(word,pword) is the association of studied word with the positive word (equivalent negative).

If the sum is positive, the word is oriented positively, and if the sum is negative, the orientation is negative. The sum absolute value indicates the orientation intensity. In order to calculate the association measure between words - A, there are several possibilities. One of them is called The Pointwise Mutual Information - SO-PMI (proposed by Church and Hanks).

$$PMI(mot_1, mot_2) = log_2 \frac{p(word_1 \& word_2)}{p(word_1)p(word_1)} \tag{4}$$

The $p(word_1 \& word_2)$ defines the probability that the two words coexist together.

Another possibility to analyze the statistical relationship between words in the corpus is the utilization of the technique: the Singular Value Decomposition (SVD).

4.4 Different approaches for the sentiment analysis

Turney's Approach

The semantic orientation of words has been elaborated first of all for the adjectives [11], [30]. The works on the subjectivity detection have revealed a high correlation between the adjective presence and the subjectivity of phrases [12]. This observation has often been considered as the proof that some adjectives are good sentiment indicators. A certain number of approaches based on the adjectives presence or polarity have been created in order to deduct the text subjectivity or polarity. One of the first approaches has been proposed by Turney [27] and can be presented in four stages:

- First of all, there is a need to make phrase segmentation (part-of-speech)
- Then, we are putting together adjectives and adverbs in series of two words
- We apply afterwards SO-PMI in order to calculate the semantic orientation of each detected series,
- Finally, we carry out a review classification as positive or negative by calculating the average of all find orientation.

Results obtained by this approach are different compared to the field: for cars= 84%, for banking documents= 80% and for cinematographic reviews= 65%. The fact that adjectives are good opinion preachers is not diminishing the other words signification. Pang et al. [21], in the polarity study of cinematographic criteria, have demonstrated that using only adjectives as characteristics gives result less relevant than using the same number of unigrams.

Pang's Approach

Pang and Lee [22] are proposing another approach for the polarity classification of cinematographic reviews. The approach is composed of two stages Figure 4. The first goal is to detect the document's parts, which are subjective. Then, they are using the same statistical classifier to detect the polarity only on subjective fragments detected previously. Instead of doing the subjectivity classification for each phrase separately, they admit that they can see a certain degree of continuity in the phrases subjectivity - a writer generally is not changing often between the fact to be subjective or objective. They give preferences in order to have proximity phrases, which have the same level of subjectivity. Every phrase in the document is then labeled as subjective or objective in the process of collective classification.

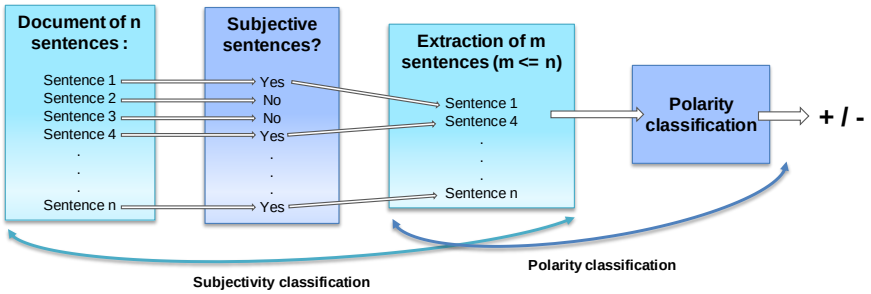

Fig. 4. Pang's approach - Utilisation of the same classification technique for the detection of subjectivity and afterwards of phrases polarity labeled as subjective.

5 Linguistic Analysis

Sentiment detection and marking can also be carried out by NLP techniques - Natural Language Processing. Information extraction consist to identify precise data of a text in natural language and to represent it in a structured form [23]. It is a documentary research, which aim to find back in the corpus a set of relevant documents regarding a question [28]. It consists in building up automatically a data bank from texts written in natural language. It is

not a matter of giving unprocessed text to use, but to give precise answers to questions that have been ask by a formula or database padding.

The extraction requires specialized lexicon and grammar. The adjustment of such resources is a long and tiresome task, which needs most of the time an expertise of the tackled field and knowledge in computing linguistic . Among this knowledge, we can mention filtration techniques of documents categorization and data extraction.

Systems of text understanding have been conceived, for most of them, as generic system of understanding, but it has been reveled not much usable in reel applications. Understanding is seen as a transduction, which transform a linear structure. It means that the text (i.e. the linear structure) is transformed in an intermediary logical-conceptual representation. The final objective is then to create inferences on those representations in order to conduct different processing, for example, answering questions.

In order to understand the whole text, there is a need to carry out the syntactic and the semantic analysis. The syntactic analysis is the largest possible because of the ambiguity problem. The semantic analysis aim to produce a structure representing, the most reliable possible, the entire sentence, with its nuances and its complexity, and then to integrate all produced structures in a textual structure. At the end, we obtain a logical-conceptual representation of the text. The semantic representation varies from one system to another.

This has driven an important number of researchers to describe natural languages in the same way as formal languages. Maurice Gross to precede, with his LADL team, the exhaustive examination of simple French phrases, in order to dispose of reliable and calculated data, on which it will be possible to conduct meticulous scientific experiences. To reach this result, each verb has been study so as to test if it verify or not syntactical proprieties as the fact to admit a completive proposition as a subject emplacement. We will see that we can not describe French with general rules. The same situation applies to all other languages. Results of this research have been encoded in matrices called lexical-grammar tables. This table shows a precise description of the syntactic behavior of each French word. The aim is to use all the resources of the lexical- grammar tables in order to obtain a system able to analyze any simple phrase structure. The sense minimal unit, according to Maurice Gross, is the sentence not the word. The principle is therefore to study the transformation that simple sentence can have. Simple sentences have been indexed via their verbs. For a verb, we can have several different usages. Thanks to syntactical proprieties, we can distinguish the usage of a verb. There are no verbs, which possess exactly the same syntactical behavior. We can not express, therefore, general rules, which could explain the language.

The text corpuses are represented by automates; in which each path correspond to a lexical analysis. Linguistic phenomena are represented by local grammar, which is translated in automates in a final stage in order to be easily confronted to the text corpuses. A local grammar [10] is a representation by automate of linguistic structures, difficult to formalize in lexical- grammar ta-

bles or in dictionaries. Local grammars represented by graphs, are describing elements that are part of the same syntactic or semantic field.

Linguistic descriptions describe in local grammar form are used for a huge variety of automatic applied processing on the text corpus. Thus, different methods of lexical disambiguates have been developed in order to carry out grammatical constraints describe with the help of this type of graph.

6 Opinion marking module

Our system possesses a modular architecture. Its principal tasks are the following: research and collect of reviews on Internet, attribution of a mark for each review and presentation of the findings. Each task is done by a specialized module [Figure 5].

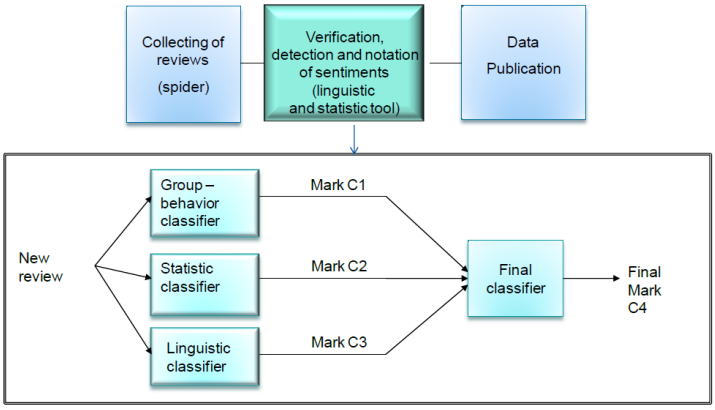

Fig. 5. General architecture of the system (the three principal modules and cinematographic reviews marking)

First of all, for the opinion marking part, we have developed three different methods for the attribution of a mark to a review:

- The group behavior classifier [section 6.1]
- The statistical classifier [section 6.2]
- The linguistic classifier [section 6.3]

Those measures are based on different approaches of document classification. Secondly, we have developed, for each method, a classifier , which assign separately the mark [9], [7]. We have obtained, therefore, three marks for each review, which can be different. We have used, finally, another classifier , which assign the final mark for the review, based only on the three marks attributed previously in the classification process [8]. For the calculation of the final

mark, we have used the values of the three marks previously attributed and their probabilities.

On a research point of view, the most important part of the system conceived is the opinion marking module.

6.1 The group behavior classifier

In this section, we present the classifier used for the opinion marking. The general approach is based on the verification that reviews, having the same associated mark, have common characteristics. Then, we determine a reviews behavior, for those having the same mark. We determine therefore, the general behavior of each review group (5 groups corresponding to five different opinion marks). We have a huge number of reviews already marked. We have gathered together all the reviews according to their mark. We obtain, then, five different groups of music marks. Afterwards, we have tried to determine typical characteristics of each group. We have defined all parameters, which can characterize the group behavior such as:

- Characteristic words,
- Characteristic expressions,
- The phrase length,
- The opinion size,
- The frequency of several words repetition,
- The negation,
- The number of punctuation signs (!, ;), ?)

The choice of criteria that we have kept for the analysis of the group behavior has been done in an empirical way. Fist of all, by analyzing the reviews corpus, we have defined criteria that seems interesting and that could determine group behavior. Then, we have tested those criteria on a training base containing a thousand of reviews. If results showed differences between groups, we considered those criteria as valid criteria for our research work. In this approach, we present the statistical study on linguistic data. The training base has been used for the review analysis, of those having the same mark, in order to find characteristics, which determine the behavior of each group. Each approach used in our research is based on different characteristics, in order not to repeat them in the classification process. However, we have borrowed semantic classes from the linguistic approach for the creation of the words list characteristics. The utilization of those data is different in those two groups. After having select criteria that characterize mark groups, we have analyzed the corpus in order to obtain statistical results. Results show huge differences between the characteristics of those groups. The creation of the global behavior of each groups, enable to determine the group in which a new review is. We have calculated for new reviews, the distance between its characteristics and those of the groups.

6.2 The statistical classifier

In this section, we propose a general approach used in the sentiment analysis. We use this method to compare results of our approaches with the same training base. The way to carry out a classification is to find a characteristic of each category and to associate a belonging function. Among known methods, we can mention Bayes' classifiers and the SVM method. We have obtained better results for the classifier of Nave Bayes, we are going therefore to based ourselves on this classifier . In our research work, we have used this classifier first of all to determine the subjectivity or objectivity of phrases, then in order to attribute a mark to subjective phrases of the review. The general process needs the preparation of training base for two classifiers to attribute a mark. The intermediate stages are the followings:

- Preprocessing and lemmatization,
- Vectorization and calculation of complete index,
- Constitution of training base for each classifier ,
- Reduction of the index dedicate to the classifier ,
- Addition of synonyms,
- Classification of texts

We are using, for the attribution of a mark to the sentiment of the review via a statistical approach, two classifiers: a first one to filter the objective and the subjective phrases and a second one to mark the review. The marking is done only on subjective phrases. Those classifiers rely on a vectorial representation of the text of the training base. This vectorial representation needs in a first time a linguistic preprocessing for the segmentation of the phrase, for the lemmatization and for the suppression of all words, which has no impact on the sense of the document. This preprocessing has been carried out for the linguistic classifier .

We carry out the preprocessing thanks to the application Unitex. We are already disposing of linguistic resources prepared for this task as, for example, the grammar of the phrase segmentation or dictionaries. Then, we take off term with no sense, such as defined or undefined articles or prepositions. We can conduct this task because those grammatical elements have a low impact on the text sense as, for example, on the opinion described in reviews, contrary to adverbs, which give a high contribution to value judgment. Afterwards, on a training corpus, we calculate the dimension of the vectorial space of the text representation in order to carry out all lemma enumeration - the entire index. Each document is then represented by a vector, which contains the number of occurrences of each lemma present in the document. Every document of the training base is represented by a vector, which dimension corresponds to the whole index and components are occurrences frequencies of the index units in the document. Therefore, at this stage of the process, texts are seen as a set of phrases. Now, each phrase is labeled according to the construction of classifiers (the subjective classifier and the marking classifier).

Labels correspond to subjective phrases (PS) or objective ones (PO) and the estimating mark attributed to those phrases (N from 1 to 5). A phrase j of the document i is marked as followed:

$$\mathbf{V_{D_iP_j}} = (f_{D_iP_j1}, ...f_{D_iP_jk}, ...f_{D_iP_j|D|}, PS/PO, N) \tag{5}$$

Where $f_{D_iP_jk}$ represents the occurrences number of the lemma k in the phrase j of the document i. The stage of the labeling was based on the reviews' marks of the training base and subjective phrases have been labeled manually. This is how we have built the set of training necessary to the determination of classifiers of subjectivity and of sentiment marking.

The last stage of the vectorial representation of the document corpus is the reduction of the entire index dedicate to the classifier. The reduction of the complete index consists in eliminating from the vectorial space of the training base, vectors, which have many components always null. This task enables us to eliminate the noise in the classifier calculation [2]. We have used the method of mutual information associated to each vectorial space dimension. In our works, we have used two classifiers: the classification based on Bayes' model and the classification using SVM. The two methods have been tested and the best results (F-score) have been obtained by the Bayes' classifiers. It is, as a result, Bayes' classifier who was used in the system. In the process of the statistical classification, we have at first classified subjective phrases and then we have attributed a mark.

Interesting phrases to carry out the opinion marking are subjective phrases because there are the only ones which contains the author point of view. For this reason, we have first of all carried out the filtration of subjective phrases. The diagram, which represents those tasks, is shown in the Figure 6.

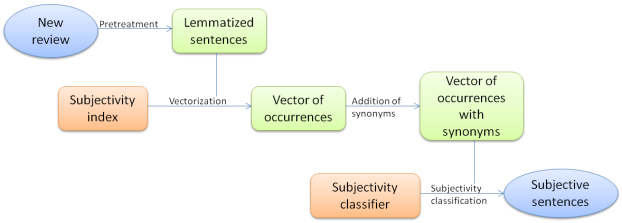

Fig. 6. Subjectivity classification - the classification steps

The process presented enables to filter only subjective phrases, those expressing an opinion. The different stages are as follow:

- The preprocessing consists in carrying out the phrase segmentation, the lemmatization and the elimination in our research of words without sense.
- The vectorization consists in putting all phrases in the form of vector of occurrences and to reduce the complete index.

- The addition of synonym consists to add terms (synonyms) in the vector of occurrences thanks to the linguistic analysis.
- The subjectivity classification consists in gathering together phrases in subjective or objective phrases. The classification is based on Bayes' theorem. For the rest of the classification (marking), we keep only subjective phrases.

After carrying out the subjectivity classification, we only keep subjective phrases. We conduct a classification in order to be able to attribute a mark to those phrases of each analyzed review. The diagram representing those tasks is presented in Figure 7. The process presented enables to attribute a mark to phrases classified in the subjective phrases. The marking varies between 1 to 5. The stages are the following ones:

- The vectorization and the reduction of the complete index dedicated to the classification of the marking
- The addition of synonyms
- The marking classification, which consists in putting together phrases according to the sentiment intensity. Marks are between 1 and 5.

At this stage of the process, we obtain marks associated to every subjective phrase. The global mark of a review of the statistical classification is the arithmetical average of all the phrases of this review.

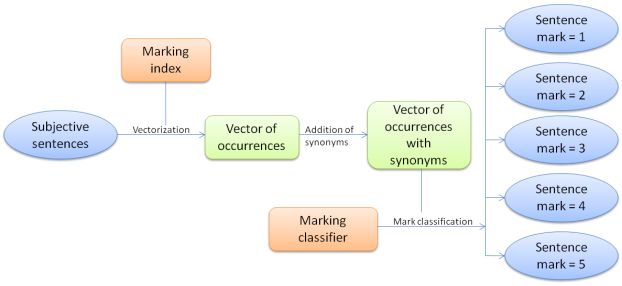

Fig. 7. Subjectivity classification - the classification steps

6.3 The linguistic classifier

We carry out reviews marking on a scale going from 1 to 5. We have created for the linguistic approach a grammar rule for each of those groups. This grammar is based on reviews' analysis of the training base, which contains approximately 2000 phrases for each mark (the same database than for the other classifiers). For this part, we have used a linguistic processing, which demand specialized lexicon and grammar. The development of those resources is a long and tiresome task, which generally needs an expertise in the field

and knowledge linguistic information processing such as filtration techniques, documents categorization and data extraction. This part of the system has been developed with the application Unitex. We are using a linguistic analyzer Unitex to conduct a preprocessing and a lemmatization of words and finally, for the most important part of our research work, the construction of complex local grammar. The example of application is shown on Figure 8 We have introduced, in order to fragment words in different opinion intensity level, words semantic categories, which are associated to words and show polarity and intensity. We have used, in order to associate words semantic categories, a subjective dictionary named General Inquirer Dictionary.

The principal goal of the linguistic classifier is the attribution of a mark according to sentiment described in the review. The marking is done phrase by phrase. The reviews' study of the training base has been carried out in the aim of creating grammar rules for each mark (in this case, the mark is between 1 and 5). Five grammars has been therefore create, one for each mark. Each grammar contains a huge number of rules taken from local grammar. For each grammar, more than thirty local grammars have been created. The analysis is done phrase by phrase to attribute a mark to a new review in order to find a rule (from our rules base) corresponding to the studied phrase. At the end of this processing, we obtain phase of the new studied review with matching grammar rules. The final mark of this classification is the average of marks corresponding to general grammars.

The construction of local grammar has been carried out manually via phrases analysis of the reviews having the same associated mark. Local grammar can not be too general because this tends to add ambiguity to results. However, if the grammar is too specific and complex, the use of this grammar is indeterminate because silence grows in a significant way. Grammars have been created to detect the opinion polarity and intensity in a phrase thanks to the local grammars form, which constitute a general grammar for each marking group. Research works are based only on local grammars form. Other characteristics purely statistical like words or characteristic expressions, phrase size, words frequency, words repetition, the number of punctuation signs and so on, are not taken into account. Of course, characteristic words are in dictionaries with semantic categories and in local grammar, but this approach is a linguistic processing (grammar is necessary) not a statistical one (like the two other classifiers).

The creation of local grammar is a tiresome task. Grammars used in our system have been created in an empirical way. We have carried out in the following way: first of all, we have constructed general grammars, then we added a complexity level to the linguistic analysis and we have made tests. After those tests, we have repeated the process (addition of a complexity level). For each level, we have conducted tests and calculated the F-score. The final result of grammars rules forms have been chose in order to obtain the best result of F-score. Unfortunately, we can not be sure of the fact that our choice is the most coherent one. We have taken into account the fact that each classifier

presented in our system should have its own criteria and characteristics. It is important to mention that the linguistic classifier provide the best results. We can observe, in particular, that the precision parameter is better than the one obtained by using other approaches.

This part of the system has been conducted with the text analyzer Unitex. Unitex enables to process in real time texts of several mega-bytes for the indexation of morphosyntactic patterns, the research of hard or semi-hard expressions, the production of concordance and the statistical study of results.

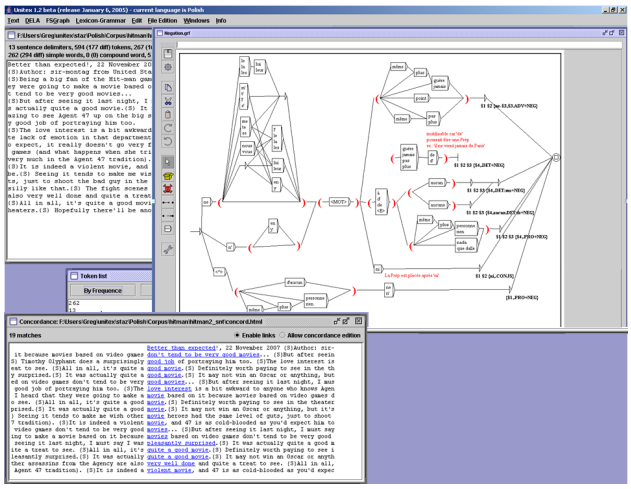

Fig. 8. Example of linguistics resources

6.4 The final classifier

Until now, we have presented three different methods to attribute a mark to a review. Thus, we obtain three different estimations (one for each classifier). The marking is carried out each time in a different way. Marks are therefore not always the same. As we are obtaining three different marks, another problem consists in conducting the final marking in order to attribute only one mark to the review. We need a final classification to obtain the final mark, which will be retransmitted to our radio. We have observed that, if we are calculating the final average obtained by the three classifiers, results are less efficient than those obtain by the linguistic classifier.

We have also observed that often a classifier in specific situations gives best results, whereas in other circumstances, it would be another one. For example [Figure 9], we have observed that often when the first classifier gives a mark equal to 2 and the last two ones give a mark of 1, the correct results is

2. As a consequence, the first classifier is determinant in this case. By implementation of neural networks for this stage and by taking into consideration each probability for each score for each classifier we improved our results for 3 to 7% depending on the class.

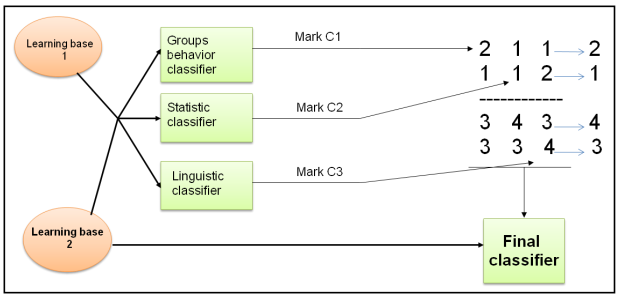

Fig. 9. Final classification- the marks behavior shows the presence of a determinant classifier in some situations

We are using, for this reason, a final classifier. For this classification we are applying a neural network. The choice of this classifier is justified by the presence of a wide reviews base, already annotated, which will be useful for the training base. Moreover, it is easy to implement those data, for it to be used in the training base. The classifier takes into account only the probability of the mark of each classifier. No other characteristics are taken in consideration. This choice is acceptable because we think that we have used all other possible characteristics in the marking process (by using the three classifiers mentioned previously) and we do not wish to repeat those characteristics in the classifications. Furthermore, the utilization of a characteristic of an opinion marking classifier in the final classification can influence the choice of this classifier.

For the entries of the final classifier, we have used marks of the previous classifiers. The marks of each classifier represented by the belonging probability of one of the five marks categories. For example, the linguistic classifier attributes the mark in the following way: the probability that the mark is:

- equal to 5 is p_5=0,6
- equal to 4 is p_4=0,2
- equal to 3 is p_3=0,1
- equal to 2 is p_2=0,1
- equal to 1 is p_1=0

We have used a neural network to determine the correlation between marks obtained by the three classifiers. We are using the neural network of multilayer perceptron (PMC) with the algorithm of retro propagation of gradient.

7 Results

We have observed for the base of cinematographic reviews that we obtain the best result with the linguistic classifier (especially for the precision). The worst results are those of the statistical classifier of Nave Bayes (the playback is correct but the precision is too low). This is demonstrating that it is necessary to carry out a deep linguistic analysis. We have observed that the best results find for the three approaches were those expressing extreme opinions [Figure 10].

Knowing the principle that it is an obligation to dispose of grammars more complex, we have demonstrate that the linguistic classifier gives better results than the statistical or the group behavior ones [Table 1].

Table 1. Classifiers results

	mark 5	mark 4	mark 3	mark 2	mark 1
Linguistic classifier	85 %	77.6 %	72.9 %	69.6 %	77.8 %
Group behavior classifier	73.8 %	71 %	70.8 %	66.1 %	68.3 %
Statistic classifier	70 %	70.7 %	66.1 %	63.3 %	73 %
Final classifier	83.1 %	81.2 %	74.5 %	72.2 %	81.4 %

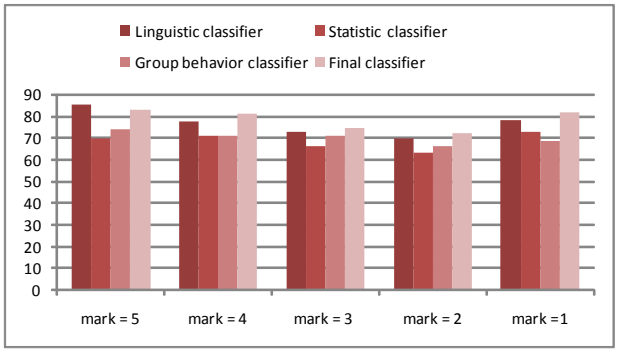

Fig. 10. Final classification- the marks behavior shows the presence of a determinant classifier in some situations

8 Conclusion

In this chapter we presented a entire system based on social network for radio application. Our radio will update a playlist depended on the votes of auditors. Users can express their selves via forums, blogs, or directly by adding a

mark to the song. In the goal of understanding the opinions written in natural language, an Opinion Mining knowledge was necessary to implement. For this reason, we presented in this chapter new approaches to automatically detect opinion from the text. The two classifications (group conduct and linguistic) have been proposed by us. Then, we have compared our approaches with the approach generally used in this field (the statistical classification, which is based on Nave Bayes' classifiers).

After carrying out tests, we can observe that we have succeeded to implement a first innovative method based on a linguistic classifier . The results obtained after this classification give us satisfaction. We can, therefore, conclude that the linguistic analysis, which is deeper, is an important research path in the field of Sentiment Analysis.

Despite the fact that the linguistic classifier enables to obtain the best results, its utilization can not be universal. Its application to a new field requires the creation of a new linguistic resource base and it is necessary to carry out the deep linguistic analysis again. Those processing are unavoidable because the language is highly dependent of the field.

References

1. Beineke, P., Hastie, T., & Vaithyanathan, S. 2004. Exploring sentiment summarization. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text, AAAI technical report SS-04-07, 2004.
2. Cover, T.M., & Thomas, J.A. 1991. Elements of Information Theory. John Wiley & sons, NY, 1991 (ISBN 0-471-06259-6).
3. Dagan, I., Karov, Y., & Roth, D. 1997. Mistakedriven learning in text categorization. 2nd Conference on Empirical Methods in Natural Language Processing, EMNLP-97, 55-63.
4. Das, S., & Chen, M. 2001. Yahoo! for Amazon : Extracting Market Sentiment from Stock Message Boards. In Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA).
5. Dave, K., Lawrence, S., & M., Pennock D. 2003. Mining the Peanut Gallery : Opinion Extraction and Semantic Classification of Product Reviews. In Proceedings of WWW, 519-528.
6. Drucker, H., Vapnik, V., & Wu, D. 1999. Automatic text categorization and its applications to text retrieval. IEEE Trans. Neural Netw., 10, 10481054.
7. Dziczkowski, G., & Wegrzyn-Wolska, K. 2007b. Rcss - rating critics support system purpose built for movies recommendation. In : Advances in Intelligent Web Mastering. Springer.
8. Dziczkowski, G., & Wegrzyn-Wolska, K. 2008a. An autonomous system designed for automatic detection and rating of film. Extraction and linguistic analysis of sentiments. In Proceedings of IEEE/WIC/ACM inter. conference on Web intelligence and intelligent agent technology, Sydney, 847–850.
9. Dziczkowski, G., & Wegrzyn-Wolska, K. 2008b. Tool of the intelligence economic : Recognition function of reviews critics. In ICSOFT 2008 Proceedings. INSTICC Press, 218–223.

10. Gross, M. 1997. The construction of local grammars. Finite-State Languauga Processing, MIT Press, 329-354.
11. Hatzivassiloglou, V., & McKeown, K. 1997. Predicting the Semantic Orientation of Adjectives. In Proc. of the Joint ACL/EACL Conference, 174-181.
12. Hatzivassiloglou, V., & Wiebe, J. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING).
13. Joachims, T. 1998. Text categorization with support vector machines : learning with many relevant features. 10th European Conference on Machine Learning, ECML-98, 137-142.
14. Joachims, T. 1999. Transductive inference for text classification using support vector machines. 16th Inter. Conf. on Machine Learning, ICML-99, 200-209.
15. Joachims, T., & Sebastiani, F. 2002. Guest editors introduction to the special issue on automated text categorization. J. Intell. Inform. Syst., 18, 103-105.
16. Knight, K. 1999. Mining online tex. Commun. ACM 42, 1, 58-61.
17. Lewis, D. D., & Gale, W. A. 1994. A sequential algorithm for training text classifiers. 17th ACM International Conference on Research and Development in Information Retrieval, SIGIR-94, 3-12.
18. Lewis, D. D., & Haues, P. J. 1994. Guest editorial for the special issue on text categorization. In : ACM Trans. Inform. Syst.
19. Mediametrie, $www.mediametrie.fr/new.php?rubrique = rad\&news_id = 229$. 2008.
20. Mitchell, T.M. 1996. Machine Learning. McGraw Hill.
21. Pang, B., Lee, L., & Vaithyanathan, S. 2002. Thumbs up ? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 79-86.
22. 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 (ACL), 271-278.
23. Pazienza, M. T. 1997. Information Extraction. In : Lecture Notes in Computer Science Vol. 1299.
24. Porter, W. A. 1980. Synthesis of polynomic systems. j-SIAM-J-MATH-ANA, 11, 308-315.
25. Schmid, H. 1994. Part-of-Speech Tagging with Neural Network. 15th conference on Computational linguistics, 1, 172-176.
26. Shannon, C. 1948. A mathematical theory of communication. Bell System Technical Journal, 27, Bell System Technical Journal.
27. Turney, P. 2002. Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the Association for Computational Linguistics (ACL), 417-424.
28. Voorhess, E.M. 1999. Natural language processing and information retrieval. Information extraction, toward scalable, adaptable systems, Lecture Notes in Computer Science, 32-48.
29. Wang, Y., Hodges, J., & Tang, B. 2005. Classification of Web Documents using Naive Bayes Method. In Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, 560-564.
30. Whitelaw, C., N., Garg, & Argamon, S. 2005. Using appraisal groups for sentiment analysis. In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM), 625-631.

31. Wiebe, J.M., Wilson, T., Bruce, R., Bell, M., & Martin, M. 2004. Learning Subjective Language. Computational Linguistics, 30(3), 277-308.
32. Wiebe, J, & Mihalcea, R. 2006. Word Sense And Subjectivity. In Proceedings of the 21st Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL), 1065–1072.
33. Wilson, t., Wiebe, j., & Hwa, r. Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In Proc. of AAAI, 761-769. Ext. version in Computational Intelligence 22(2, Special Issue on Sentiment Analysis),73-99, 2006.
34. Wilson, T., Wiebe, J., & Hoffmann, P. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 347-354.
35. Yang, Y., & Liu, X. 1999. A re-examination of text categorization methods. 22nd ACM International Conference on Research and Development in Information Retrieval, SIGIR-99, 42-49.

Index