Sentiment Analysis: A tool for Rating Attribution to Content in Recommender Systems

Antonis Koukourikos[1], Giannis Stoitsis[2,3], Pythagoras Karampiperis[1]

[1]Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research “Demokritos”Agia Paraskevi Attikis, P.O.Box 60228, 15310 Athens, Greece
[2]Agro-Know Technologies, Grammou 17 Str., Vrilissia Attikis, 15235 Athens, Greece
[3]Universidad de Alcalá, Pza. San Diego, s/n - 28801 Alcalá de Henares, Madrid, Spain

Abstract. Collaborative filtering techniques are commonly used in social networking environments for proposing user connections or interesting shared resources. While metrics based on access patterns and user behavior produce interesting results, they do not take into account qualitative information, i.e. the actual opinion of a user that used the resource and whether or not he would propose it for use to other users. This is of particular importance on educational repositories, where the users present significant deviations in goals, needs, interests and expertise level. In this paper, we propose the introduction of sentiment analysis techniques on user comments regarding an educational resource in order to extract the opinion of a user for the quality of the latter and take into account its quality as perceived by the community before proposing the resource to another user.

Keywords: Recommender Systems, Educational Repositories, Sentiments Analysis, Qualitative Analysis

1 Introduction

Recommender Systems are of particular importance within social environments, where users share access to a common set of resources. The variability of crucial user characteristics, like their background, their special interests, their degree of expertise, pose interesting issues in terms of proposing a resource that is interesting, useful and comprehensible to a particular user.

Collaborative filtering approaches based on explicitly given user ratings do not always reflect the differentiation between the various criteria that apply to a resource and the weight that the users give to each criterion. On the other hand, techniques that examine access patterns may suffer from the appearance of stigmergy phenomena. The visibility of a resource, or even more elaborate features like the time spent in a resource, the amount of downloads etc. are not directly connected to its quality or suitability. Hence, the examination of access and use patterns can lead to poor recommendation that will be further propagated due to the users continuing to follow previously defined paths within the repository of available content.

In this context, we propose the exploitation of user generated comments on the resources of a repository of educational content in order to deal with the lack of explicit ratings and discover qualitative information related to a specific resource and the impressions it left to the users that accessed it. To this end, we applied sentiment analysis to comments on educational content and examined the accuracy of the results and the degree to which they reflect user satisfaction.

The rest of the paper is structured as follows: we provide a brief review of the sentiment analysis in Section 2. We present the four algorithms that we aim to implement and examine for the Organic.Edunet recommendation system in Section 3. Section 4 describes the experimental setup and the results for the first of the proposed approaches. We conclude with our conclusions so far and report on the intended next steps.

2 Related Work

Recommender systems, particularly using collaborative filtering techniques, aim to predict the preferences of an individual (user/ customer) and provide suggestions of further resources or entities (other users of the same system, resources, products) that are likely to be of interest. The usage of recommender systems is widely spread in e-commerce environments [1] but the general principle is applicable to multiple and diverse environments. In the case of TEL, multiple solutions have been proposed and examined [2, 3]. Due to the particularities of the domain, some of the most common algorithms for collaborative filtering have been shown to struggle in the setting of a learning object repository [4, 5]. As mentioned, the presented techniques are examined in order to be incorporated in a recommender system over a social platform that provides access to educational content. Linguistic techniques, such as sentiment analysis, can be of use for alleviating some of the drawbacks of traditional algorithms in terms of differentiating users belonging in different audiences (e.g teachers from students) and bypassing the need for explicit ratings (via a star system).

Sentiment analysis regards extracting opinion from texts and classifying it into positive, negative or neutral valence [6]. Work on the field focuses on two general directions; lexical approaches and solutions using supervised machine learning techniques.

Lexical approaches rely on the creation of appropriate dictionaries. The terms present in the dictionary are tagged with respect to their polarity. Given an input text, the presence of dictionary terms is examined and the overall sentiment of the text is computed based on the existence of "positive" and "negative" terms within it. Despite its simplicity, the lexical approach has produced results significant better than "coin-toss" [7, 8, 9]. The way of constructing the lexica that are used for sentiment analysis is the subject of several works. In [10] and [11] the lexicons comprised solely adjective terms.
The usage of pivot words (like "good" and "bad") and their association with the target words is also a frequently met approach. In [9] and [12], the minimum path between each target word and the pivot terms in the WordNet hierarchy was calculated in order to determine the polarity of the term and its inclusion in the dictionary. In [8], the authors executed search queries with the conjunction of the pivot words and the target word given as input. The query that returned the most hits determined the polarity of the given word.

Machine learning techniques focus on the selection of feature vectors and the provision of tagged corpora to a classifier, which will be used for analysing untagged corpora. The most frequent routes for choosing the feature vectors are the inclusion of unigrams or n-grams, counting the number of positive/ negative words, the length of the document etc. The classifiers are usually implemented as a Naive Bayes classifiers or as Support Vector Machines [9, 13]. Their accuracy is dependent on the selection of the aforementioned feature vectors, ranging in the same space as the lexical approaches (63%-82%).

3 Algorithms under Analysis

For our experiments, we aim to examine the following sentiment analysis algorithms and evaluate their performance in order to deploy the most suitable for a repository of educational content.

The fact that we are dealing with user generated content drives us to take into account its unstructured nature and the potential unbalanced distribution it may present. This gives rise to the fact that our training set may be unbalanced and therefore learning may not be able to cope with such diversity in the number of instances per class. Hence, these properties require simulating sentiment representations onto which the input text will be mapped, since sentiment prediction calls for predefined knowledge. Therefore, we focus on lexical approaches for capturing the polarity expressed in a comment. Specifically, we produced implementations of the following algorithms.

3.1 Affective Terms Frequency

Rather small documents or text chunks that carry a certain kind of sentiment polarity have been found to present that valence throughout the text, or in most parts of it. For example, in tweets we observe cases like: "Don't you just love this camera? It's great!" In such a piece of discourse, the probability of tracing negative terms is rather low.

This observation gives rise to determining the affective term frequency that appears in cases as the above mentioned. In order to capture the overall sentiment expressed in such inputs, we proceed as follows:

The algorithm receives as input the text to be processed and two lists of affective terms, one of positive and one of negative valence.

For every word of the text to be processed, we examine whether it is mapped on either of the two lists [14]. If it matches an entry of in either of them, a corresponding value gets incremented by 1. After having traversed the whole text, we compare the two sums and the one with the highest value is considered to be the dominant one and the respective valence is attributed to the input text [15]. If they are equal or no sentiment is detected, the text is considered to carry neutral sentiment.

In specific, we calculate:

$$\boldsymbol{\Delta} = \sum_{i=0}^{n-1} \boldsymbol{x_i} - \sum_{j=0}^{n-1} \boldsymbol{y_j} \ (1)$$

where, Δ is the absolute difference between the positive and negative sums, x_i is each instance in the vector representing the positive valence, y_j is each instance in the vector representing the negative valence and n is the number of words the input text may contain.

If $\Delta > 0$, the sentiment is positive, if $\Delta < 0$, the sentiment is negative, else the sentiment is neutral.

3.2 Weighted Affective and Domain Term Frequency

Sentiment attribution varies according to the domain(s) the text to be examined belongs to. For example, the term "small" is considered to be negative, due to its connotation with insufficiency. However, in the domain of computer hardware and mobile devices, its valence is shifted and it's mostly attached with positive opinions. For instance, in a product review or in a forum, we mostly come across statements like: "The iPad 2 is amazing! It's so small and light I can take it with me everywhere". In order to be able to identify such assertions, we may need to add to every affective term a frequency value that will determine how positive, negative or neutral it is [16].

As a deduction, the same principle applies for all the text terms (excluding stop words). The reason for such a precaution is that specific terms that don't appear to influence text valence, in general, change their sentiment in regard to a specific context. For instance, the verb "watch" doesn't seem to present any specific sentiment. However, in movie reviews, we come across comments like the following: "You should definitely watch this movie". In this example, a positive opinion can be detected neither in a word, nor in an idiom or irony or any other discourse schema. Yet, the sentence bears it.

One may argue that this doesn't guarantee any credibility, since another user comment of e.g. IMDB could be: "You should definitely not watch this movie". Nevertheless, the reason the sentiment expressed in the second comment is opposite to the first one is the presence of the valence shifter: "not" and not of an affective term or discourse schema denoting negative valence. This would be the case if the two comments would be the following respectively: "I enjoyed this movie!" versus "I did not enjoy this movie". As we can tell, in either case, the sentiment of the affective term is reversed, irrespectively of the term belonging to a list of affective terms, i.e. the term "enjoy", or not belonging to a list of affective terms, i.e. the term "watch".

The algorithm receives as input the text to be processed and three hash tables of affective terms and their frequency, one of positive, one of negative and one of neutral valence.

The hash-table lists of affective terms are constructed as follows: A set of domain specific terms is built [17, 18]. Depending on the domain, the corpus may consist of product reviews, critiques on results of intellectual effort (music, movies) or more formal documents like questionnaires, review forms etc. Human annotators examine the polarity of this content and the corpus is partitioned in three sub-corpora, one consisting of positive, one of negative and one of neutral terms. For every word in the sub-corpus of positive annotated texts, we attribute its frequency in this specific corpus. The same procedure takes place for the other two corpora. Sub-sampling may have taken place where necessary.

As a result, if a term appears in all three sub-corpora, it receives three values, one attributed to each of them respectively. If it appears in two or in one it receives the score it has been assigned within this corpus. The reason why the sum of positive, negative and neutral doesn't sum up to 1 is that, if, for example, word "nice" appears only in positive terms, it would be assigned weight value 1. At the same time, if, for example, word "excellent" again appears only in positive terms, it would be assigned weight value 1. Subsequently, such an attribution would oppose the current approach, aiming at measuring sentiment as per weighted frequencies. Such weighted frequencies get enhanced by predefined lists.

For positive, negative and neutral valence, three hash tables are created, containing the positive, the negative and the neutral lists respectively. For every word of the text to be processed, we examine whether it is mapped on either of the three hash tables. If a key in the hash table comprising the positive terms maps on the word in question, its value is added in a vector keeping the positive term frequencies. If the hash table comprising the negative terms also contains this word, its value is added in a vector keeping the negative term frequencies and so on. After the whole document has been traversed, the values of each of the three vectors are summed up, the three sums are compared and the one with the highest value designates the sentiment of the respective post.

In specific, we calculate:

$$\boldsymbol{\Delta} = \sum_{i=0}^{n-1} \boldsymbol{x_i} \cdot \boldsymbol{w_i} - \sum_{j=0}^{n-1} \boldsymbol{y_j} \cdot \boldsymbol{w_j} \ (2)$$

where, Δ is the difference between the positive and negative sums, x_i is each instance in the vector representing the positive valence of weight w_i, w_i is the value of the positive weight of instance x, y is each instance in the vector representing the negative valence of weight w_j and w_j is the value of the negative weight of instance y_i.

If $\Delta = 0$, the sentiment is neutral.

Else, if $\Delta > 0$, we calculate:

$$\boldsymbol{\Delta_1} = \sum_{i=0}^{n-1} \boldsymbol{x_i} \cdot \boldsymbol{w_i} - \sum_{k=0}^{n-1} \boldsymbol{z_k} \cdot \boldsymbol{w_k} (3)$$

where Δ_1 is the absolute difference between the positive and neutral sums, z_k is each instance in the vector representing the neutral valence of weight w_k and w_k is the value of the neutral weight of instance z_k. If $\Delta_1 = 0$, the sentiment is neutral, else it is positive.

Else, we calculate:

$$\boldsymbol{\Delta_2} = \sum_{\mathrm{j}=0}^{n-1} y_j \cdot \boldsymbol{w_j} - \sum_{k=0}^{n-1} \boldsymbol{z_k} \cdot \boldsymbol{w_k} (4)$$

where Δ_2 is the absolute difference between the negative and neutral sums. If $\Delta_2 = 0$, the sentiment is neutral, else it is negative.

3.3 Distance between Affective and Sentiment Targeted Terms

When we are called to tackle a more specific problem, e.g. a more targeted question in a more concrete domain, it is required that our processing is also more focused on the object, i.e. the entity or entities that represent it, and/or the domain towards which opinion is expressed. This way, we aim at capturing the entity characteristics that affect sentiment rendering. For example, we may come across a comment like: "I am discontent by the book I had read and I've found it rather useless, but this video is really great." The previous two approaches would classify this comment as being negative as regards the video, while the writer expresses positive opinion about it.

Moreover, comments, in specific, comprise in a lot of cases unstructured chunks and other abnormalities, such as emoticons, artificial words, e.g. "yeeeeees", punctuation mistakes, e.g. "What is that?????" and many more syntactic and grammatical deviations from standard language.

In consideration of such differentiations, we count the distance between the affective terms and the terms that are involved in the representation of the entity towards which sentiment is expressed. If sentiment is expressed towards more than one entity or entity representations, then all distances are counted recursively.

The algorithm receives as input the text to be processed, two lists of affective terms, one of positive and one of negative valence and the entity/entities towards which sentiment is expressed. In the rest of the paper we will also refer to these latter terms as "keyword(s)" for simplicity reasons.

The position of the entity towards which sentiment is expressed is tracked in the text to be processed. The words of the document to be examined are mapped onto the affective terms of the input lists. If the document contains an affective term, then its position in the text is tracked as well and we calculate the distance that separates them.

To be more specific, we detect whether the affective term precedes or follows the entity in question. After having located these two points in the text, we calculate the distance between them, in the substring that separates them [19]. This is based on word count versus character count, because in this approach words are considered to be autonomous semantic meaningful units, unlike alphanumeric strings, regarded as self-contained units in graph-based approaches.

The above mentioned procedure takes place for all affective terms. In particular, for every positive term that appears in the input text, its distance from the entity in question is counted. After all positive terms have been checked, the smallest distance is kept to be compared with the respective smallest distance between the negative terms and the entity in question. If the two values equal to zero, or are equal, neutral sentiment is attributed. Otherwise, the post receives the sentiment represented by the smaller of the two values. If we have more than one entity representation, the same procedure is applied and again the shortest distance is taken as representing the sentiment of the writer.

In specific, we calculate:

$$\Delta = x - y(5)$$

where Δ is the difference of the equation, denoting which of the two scores is higher, x is the minimum distance between the positive term and the key word and y is the minimum distance between the negative term and the keyword. If $\Delta > 0$, the sentiment is positive, else if $\Delta < 0$, the sentiment is negative, else, the sentiment is neutral.

3.4 Dependencies between Affective and Sentiment Targeted Terms

More formal documents tend to present a more consistent and accurate syntactic and grammatical structure, hence more concrete and concise textual forms. This characteristic restricts the number of alternatives we may have in expressing a certain meaning and therefore facilitates us to capture it. As a result, the better we are able to represent this structure, the closer we get in capturing the semantics it pertains [20].

In our approach, we are interested in detecting the syntactic dependencies between the keywords and the affective targeted term(s). In particular, when the sentiment analysis algorithm accepts a text as input, it accepts it attached to certain categories and/or the description of the educational material in question. As a consequence, our goal is to track the sentiment of the writer in relation to the material we are examining. For this reason, we process every sentence of the particular comment so as to identify whether a reference of the material is attached to an affective term. Syntactic Parsers provide the necessary tools to analyze the input text. An example is illustrated in Figure 1, where in the second line we can tell that, when referring to the relationship of two words we refer to words: "trustworthy, China" and the kind of relation that binds them is: "nsubj", namely: "China" is the subject of a verb and "trustworthy" is the attribute of the same verb.

```
currentSentence: China is trustworthy.
nsubj(trustworthy-3,china-1)
cop(trustworthy-3,is-2)
```

Fig. 1. Exemplary output of the Stanford Parser

The algorithm, thus, accepts as input the text to be processed, two list of affective terms (positive and negative), a list containing the keywords that define the text's context, a list of reporting verbs defining assertions, a list of verbs differentiating indirect speech to counterfactuals and a list of stop words.
The basic linguistic processing steps of sentence splitting and tokenization are performed, before obtaining the parse tree for each of the resulting sentences. The tokens are lemmatized in order to be mapped to the words included in the aforementioned lists e.g. the term "loved" should be lemmatized to "love", so as to be mapped on the corresponding entry of the list containing the positive terms. Next, we extract the dependencies between the input text lemmas and the words that are contained in our list. This procedure takes place in order to acquire those types of dependencies that will infer the sentiment polarity we will attribute, as we've shown in Figure 1.

After having tracked a keyword or keywords, we try to detect affective terms. If no such terms are tracked, then the sentence is considered as carrying neutral sentiment. This value is kept in a counter whose value is increased by 1 every time a sentence of neutral polarity is met. At the end of each post processing, all neutral values are being accumulated and point out a neutral sum, to be compared with the positive and negative ones.

On the other hand, if a sentiment-bearing word is tracked, we try to identify whether this sentiment word renders sentiment to the word/words describing the material in question. If the examined text reports the beliefs of another person, the sentence being examined is considered neutral. To identify such cases, we use the respective lists given as input to the algorithm. If the sentence contains a verb also found in the assertions or counterfactuals lists, the process is stopped, the sentence is appointed with a neutral value and the analysis continues for the next sentence. In the case of the existence of verbs denoting counterfactual, the list is employed taking into consideration that we contemplate at segregating secondary if-clauses that are dependent from a question verb, that is when the main clause they depend on regards indirect speech, versus if-clauses that don't depend on question verbs, that is when the main clause they depend on regards counterfactuals [21].

Otherwise, we investigate the existence of valence shifters within the examined sentence. At this moment, we take into account negations and comparisons. In the first case, if a word that discloses negation is present (e.g. "no", "not", etc.) and it is syntactically associated with the found affective term, the latter term is considered to carry the opposite polarity value. In the case of comparisons, the affective term pertains to both the compared entities. We distinguish two general cases:

- One entity accepts the actual valence of the affective word and the other one the opposite. In specific, the valence to be accredited is decided in reference with the syntactic relationship between the keyword term under examination and the word that discloses comparison (e.g. "than").
- Both entities accept the valence of the affective word. Specially, the valence to be accredited is decided in reference with the syntactic relationship between the keyword term under examination and the word that discloses comparison (e.g. "as").

We first eliminate stop words from the keyword list. Therefore, a new set of keywords is created. For every word of the text to be processed, we examine whether it is mapped on this new set and on the other four lists. For every sentence of the input text, if an entry of the lists in the reporting verbs is met and the dependency that binds it to the text's opinion holder is of subject type, neutral sentiment is attributed. Else, for every word of the text to be processed, we examine whether it is mapped on either of the two lists containing the affective terms. If they match an entry in either of them, if no negation or comparison dependencies are met, a corresponding value gets incremented by 1, else, the reverse one.

After having traversed the whole text, we compare the two sums and the one with the highest value is considered to be the dominant one and the respective valence is attributed to the input text. If they are equal or no sentiment is detected, the text is considered to carry neutral sentiment.

4 Results

A set of experiments have taken place so as to evaluate the performance of each algorithm. At this moment, we have completed and present here the results for the first of the presented algorithms.
Given the fact that our task is a classification one, standard classification metrics from the literature have been used. In specific, we try to detect the precision, recall and accuracy values obtained from the above described input data sets.

We wanted to detect opinion in three classes, i.e., positive, negative and neutral. So, precision will show us for each of the positive class how many of the positive instances found are indeed positive; recall will show how many of the positive instances have been found out of the total number of the positive instances that should have been found are indeed positive. The same measures will be given for the other two classes; finally, accuracy will show for each data set how many instances were correctly classified as far as all three classes are concerned.

For an initial corpus of user generated reviews, we used content from the Merlot[1] repository. Merlot is an online repository distributing free access to resources for learning and online teaching. It provides learning material of higher education aiming at promoting access to scientific data and as a result to their manipulation and exploitation by research communities. Reaching its instructional objectives necessitates ensuring that the quality of its content is of high standards. It, therefore, accredits reviews and peer reviews, attending on continuously enhancing their quantity and quality. Our system aims at enabling this procedure by proposing a way of evaluating automatically opinions expressed for the learning materials and thus contributing to enabling the community accessing valuable data and promoting its scientific goals.

[1] www.merlot.org/

Within Merlot, we are interested in the user comments and the expert reviews associated with each educational resource. To be more specific, users and community experts have expressed their opinion in respect of its quality, its orientation and the degree to which it complies with helping the user exploit its potentials. We refer to the former category as “user comments” and to the latter as “expert reviews”. The expert reviews provide an evaluation for three distinct subcategories, namely (a) Content quality, (b) potential effectiveness as a teaching tool and (c) ease of use for both students and faculty.

For each category of the corpus we have performed two experiments, as provided by the two sets of lists respectively. Our first category regards the processing of the 6792 user comments stored in the Merlot repository. These comments have been considered as attributing positive opinion with respect to a research material if they have been rated with 5 or 4 stars, neutral if they have been attributed 3 and otherwise negative.

As peer reviewers state their opinions with respect to strengths and concerns in each of the aforementioned subcategories, the neutral class is empty in this context. To be more specific, for each subcategory, we have tested our system’s performance again with both sets of lists in the 626 texts carrying sentiment.

4.1 Construction of the Lists of affective Terms

For the experiments conducted thus far, two sets of lists from the literature have been tested as input, both of which contain positive and negative terms. No list of neutral terms has been taken into consideration, since literature doesn’t provide such lists.

The first set of lists is provided by [22] and we will refer to it as the “ANEW” subset. The second is derived from SentiWordNet [23].

Namely, SentiWordNet is a lexical resource for opinion mining. It assigns to each synset (synonym set) of WordNet three sentiment scores, each representing respectively: positivity, negativity, objectivity. In specific, according to WordNet, a synset or synonym set is defined as a set of one or more synonyms that are interchangeable in some context without changing the truth value of the proposition in which they are embedded.

The values of positivity, negativity and objectivity follow the rule:

$$positivity + negativity + objectivity = 1 (6)$$

where,:

- positivity describes how positive the terms contained in the synset are,
- negativity describes how negative the terms contained in the synset are and
- objectivity describes how neutral the terms contained in the synset are.

Our goal was to extract two lists of words, positives and negatives. Due to the fact that both positivity and negativity values are assigned to a word we needed to make sure the word was clearly biased. So, each of the three classes can take values from 0 to 1 and they are complementary, as deduced by the formula.

The rule we've used for extracting the lists is:

$$\frac{positivity}{negativity} \geq 0.7 \;\&\&\; objectivity \geq 0.2 (7)$$

This way we check that the word has a positivity or negativity value above 70% and from the rest of the percentage, at least 20% goes to objectivity leaving only 10% max for the opposite sentiment.

By applying the check defined in (7) we make sure there is a clear bias towards the positivity or negativity and the rest is assigned to objectivity.

Finally, we have created a subset of lists from the two above mentioned subsets, i.e. the ANEW and the SentiWordNet ones. To be more specific, two hash tables have been created one containing the positive ANEW terms and the other the positive SentiWordNet terms. If a key of the first table wasn’t also a key entry in the second one, it was added in the new list. Having applied the same de-duplication procedure for the negative terms, we obtained two new lists, containing all terms of the first and the second list with unique entries.

4.2 Results for the Affective Term Frequency Algorithm

The respective results of each subcategory are presented in the following tables.

Tables 1 and 2 show the precision and recall achieved by the current system version for user comments and experts reviews respectively. What is of interest is that the User Comments present very high accuracy in the positive class, unlike the negative one. The reason for such results is the unbalanced distribution of instances per class in the specific input set. Moreover, we can tell that, when prior sentiment knowledge is received as input via the ANEW lists, precision and mostly recall is higher than when SentiWordNet or Mixed lists are adopted.

Table 1. Precision and Recall for User Comments

List	Positive		Negative		Neutral	
	Precision	Recall	Precision	Recall	Precision	Recall
ANEW	0.999	0.823	0.0	0.0	0.031	1.0
SentiWN	0.995	0.390	0.0	0.0	0.031	1.0
Both	0.996	0.740	0.0	0.0	0.010	0.242

Table 2. Precision and Recall for Expert Reviews

Subcategory	List	Positive		Negative	
		Precision	Recall	Precision	Recall
Content Quality	ANEW	0.737	0.940	0.930	0.353
	SentiWN	0.660	0.310	0.392	0.170
	Both	0.793	0.852	0.650	0.314
Effectiveness	ANEW	0.710	0.900	0.860	0.400
	SentiWN	0.704	0.458	0.707	0.220
	Both	0.721	0.853	0.643	0.371
Ease of Use	ANEW	0.860	0.844	0.864	0.270
	SentiWN	0.565	0.240	0.350	0.200
	Both	0.740	0.700	0.591	0.260

Table 3 shows the overall accuracy of the module as presented in every subcategory. Here, again, we can notice the higher values obtain by ANEW lists, followed by the Mixed ones. Furthermore, it's worth mentioning that precision and recall figures of the positive and negative classes are fairly higher than the accuracy of the overall system. The reason for such a difference lies in the fact that our input didn't include neutral class instances.

Table 3. Accuracy achieved

Input Type		Lists	Accuracy
User Comments		ANEW	0.823
		SentiWN	0.390
		Both	0.734
Expert Reviews	Content Quality	ANEW	0.550
		SentiWN	0.300
		Both	0.580
	Effectiveness	ANEW	0.534
		SentiWN	0.390
		Both	0.552
	Ease	ANEW	0.530
		SentiWN	0.262
		Both	0.471

5 Conclusions & Future Work

The preliminary results of the sentiment analysis on user comments in the context of a repository of educational resources indicated that there can be valuable qualitative information that can be added to a recommendation service and be used to adjust the perceived "rating" of a given resource by a specific user. The accuracy of the first of the examined algorithms, while satisfactory, leaves room for improvement. We expect that more elaborate techniques that introduce association of entities and contextual information will produce better results. However, it is important to note that sentiment analysis does not suffer much from domain differentiation or variability on user roles (that is, the results for expert reviews and general user comments presented similar success). An interesting remark regarding the linguistic characteristics of the examined content is that the criticism is usually employed using mild terminology, which is in contrast of user-generated reviews for products/ movies etc. This indicated the necessity of repeating the experiments with different thresholds for the restriction employed in (7), as a review considered neutral or even positive by the system is actually negative but the phrasing of the reviewer is not strong enough to provide strong indications of his/ her polarity.

Our immediate next step is to measure the performance of the remaining sentiment analysis algorithms and draw conclusions for their suitability in the context of large-scale educational repositories. Following the finalization of the sentiment analysis methodology, we intend to incorporate the results in the recommendation system for the Organic.Edunet platform in order to produce a "suitability score" for a user-resource pair or a community-resource pair. Our aim is to define this score in a way that reflects both quantitative (visits, access time, downloads) and qualitative (opinions) characteristics. The foundation of our envisioned approach is the building of a connectivity graph between the system's users and communities with respect to their profile similarity and their interests as perceived by their activity. The sentiment analysis module will be used for extracting their opinion on the overall quality of the resources they have reviewed or commented on, as well as more specific characteristics (ease of understanding, innovation) where such features can be recognized by the linguistic analysis of the reviews/ comments. The sentiment score will be incorporated in the calculation of the trust and reputation scores of the users and resources will be proposed to other members of the community based on these scores.

6 Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme, in the context of the SYNC3 (ICT-231854) project.
This paper also includes research results from work that has been funded with support of the European Commission, and more specifically the project CIP-ICT-PSP-270999 "Organic.Lingua: Demonstrating the potential of a multilingual Web portal for Sustainable Agricultural & Environmental Education" of the ICT Policy Support Programme (ICT PSP).

7 References

1. B. Sarwar, G. Karypis, J. Konstan and J. Riedl: Analysis of recommendation algorithms for e-commerce. In proceedings of the 2nd ACM Conference on Electronic Commerce, pp. 158-176, Minneapolis, MN, October 2000
2. N. Manouselis, R. Vuorikari and F. Van Assche: Collaborative Recommendation of e-Learning Resources: An Experimental Investigation, Journal of Computer Assisted Learning, 26(4), 227-242, 2010
3. N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel and R. Koper: "Recommender Systems in Technology Enhanced Learning", in Kantor P, Ricci F, Rokach L, Shapira B (eds), Recommender Systems Handbook, pp. 387-415, Springer US, 2011
4. J. L. Herlocker, J. Konstan, L.G. Terveen and J. Riedl: Evaluating Collaborative Filtering Recommender Systems. ACM Trans. Inf. Syst., Vol. 22(1), pp. 5-53, January 2004
5. M.-A. Sicilia, E. Garca-Barriocanal, S. Sanchez-Alonso and C. Cechinel: "Exploring user-based recommender results in large learning object repositories: the case of MERLOT", in Manouselis N, Drachsler H, Verbert K, Santos OC (eds) Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010), Procedia Computer Science, 1(2):2859-2864, 2010.
6. B. Pang and L. Lee: "Opinion Mining and Sentiment Analysis", Now Publishers Inc., 2008

7. K. Durant, and M. Smith: Mining Sentiment Classification from Political Web Logs.In: Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Kennedy, A., Inkpen, D.: Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters. Computational Intelligence, 110–125 (2006)
8. P. Turney: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL, Philadelphia, PA, July 2002, pp. 417–424 (2002)
9. J. Kamps, M. Marx and R.J. Mokken: Using WordNet to Measure Semantic Orientation of Adjectives. In: LREC 2004, vol. IV, pp. 1115–1118 (2004)
10. V. Hatzivassiloglou and J. Wiebe: Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In: Proceedings of the 18th International Conference on Computational Linguistics, New Brunswick, NJ (2000)
11. A. Kennedy and D. Inkpen: Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters. Computational Intelligence, 110–125 (2006)
12. A. Andreevskaia, S. Bergler and M. Urseanu: All Blogs Are Not Made Equal: Exploring Genre Differences in Sentiment Tagging of Blogs. In: International Conference on Weblogs and Social Media (ICWSM-2007), Boulder, CO (2007)
13. J. Akshay: A Framework for Modeling Influence, Opinions and Structure in Social Media. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, Vancouver, BC, July 2007, pp. 1933–1934 (2007)
14. M. Hu and B. Liu, "Mining and Summarizing Customer Reviews", in Proc. of the 10th International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 168-177, Seattle, USA, 22–25, August 2004
15. R. Feldman, B. Rosenfeld, R. Bar-Haim and M. Fresko, "The Stock Sonar — Sentiment Analysis of Stocks Based on a Hybrid Approach", in Proc. of the 23rd Innovative Applications of Artificial Intelligence Conference (IAAI 2008), pp. Chicago, USA, 13–17, July 2008
16. S. Kim and E. Hovy, "Determining the Sentiment of Opinions", in Proc. of the 20th International Conference on Computational Linguistics (COLING 2004), Article No.: 1367 pp. http://dl.acm.org/citation.cfm?id=1220555, Geneva, Switzerland, 23-27, August 2004
17. N. Godbole, M. Srinivasaiah and S. Skiena, "LargeScale Sentiment Analysis for News and s", in Proc. of the 1st International Conference on Wes and Social Media (ICWSM 2007), pp. short paper Boulder, USA, 26-28, March 2007
18. J. Kim, J. Li and J. Lee, "Discovering the Discriminative Views: Measuring TermWeights for Sentiment Analysis", in Proc. of the 47th Annual Meeting of the Joint conference of the Association for Computational Linguistics (ACL 2009) and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP 2009), pp. 253–261, Suntec, Singapore, 2-7, August 2009
19. X. Ding, B. Liu and P. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", in Proc. of the 1rst International Conference on Web Search and Web Data Mining (WSDM 2008), pp. 231-240, Palo Alto, USA, 11-12, February 2008
20. T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing", in Proc. of the 2nd International Conference on Knowledge Capture (K-CAP 2003) held in conjunction with the 2nd International Semantic Web Conference (ISWC 2003), pp. Sanibel Island, USA, 23–25, October 2003
21. R. Narayanan, B. Liu, and A. Choudhary, "Sentiment Analysis of Conditional Sentences", in Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), pp. 180–189, Singapore, Singapore, 6-7, August 2009
22. M. Bradley and P. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings", Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999
23. B. Ohana, B. Tierney, "Sentiment Classification of Reviews Using SentiWordNet", in Proc. of the 9th Information Technology and Telecommunication Conference (IT&T 2009), pp. Dublin, Ireland, 22 - 23, October 2009