Automated Intellectual Analysis of Consumers' Opinions in the Scope of Internet Marketing and Management of the International Activity in Educational Institution

Daniil Gorbushin, Dmitriy Grinchenkov, Anastasia Kolomiets and Nguyen Phuc Hau

Platov South-Russian State Polytechnic University (NPI)
Prosveshchenie Str. 132, 346400, Novocherkassk, Russia
gorinwww@gmail.com, grindv@yandex.ru, anastasia.srstu@gmail.com, phuchauptit@gmail.com

Abstract: The issue of internet marketing in the international cooperation in institutions of higher education is studied in the article. Social media marketing technologies are offered to solve the task, paying special attention to the analysis stage. It is offered to consider social networks users' messages as analytical material. Intellectual sentiment analysis is offered as a method for such analyses. Approaches to the solution of a sentiment analysis problem are considered. The perspective application methods of collecting and the intellectual analysis of information necessary for the feedback organization and ensuring high-quality educational services are described.

1   INTRODUCTION

International relations and international academic exchange are compound integral part of modern higher educational institutions functioning. International orientation plays a key role almost in all Higher Education Institutions, as well as in the common context of the higher education policy. The international research cooperation gains more and more value. Besides, graduates are often looking for the place on the world market to find themselves. Besides, increase in global budgets and target use of resources also brings need for evaluation methods of the international level.

Now this area is presented by means of the considerable set of parameters, the majority of which has gaged character, as it is realized in the existing ratings of carrying out comparisons at the international level, for example, the Shanghai rating, Times Higher Education rating, etc. Besides gaged parameters, there are also other factors which exert the considerable impact on quality of the international activity which cannot be measured quantitatively, but which can be revealed on the basis of the analysis of the opinion of consumers expressed by them, for example, on social networks.

The international experience of activity of departments of institutional researches in the leading foreign centers and universities of the USA, the Netherlands, Germany, Great Britain and other countries confirms need of an evidence-based decision making for questions of strategic scheduling and assessment of the international activity of higher education institutions (Green, 2008), (Branderburg, 2007).

Nowadays many educational organizations are interested in efficient international cooperation, at the same time a task of establishment and support of partner contacts between employees and students of universities is rather actual. A possible way to draw attention to the educational organization is to create a marketing strategy, especially on social networks and blogs, as it is one of the main important channels for modern marketing. First of all, it is due to fast-growing percentage of social networks traffic on the internet, possibility to use targeted marketing and by great number of active users.

Scientific literature review makes it possible to allocate one of the perspective directions in the field of internet marketing - Social media marketing (SMM). SMM is a process of drawing attention to a brand through social platforms. It appeared in the marketing less than ten years ago and managed to prove the high performance, with increasing number of companies of different scale resorting to its help. SMM allows acquainting target audience with a brand, drawing attention to the activity, announcing
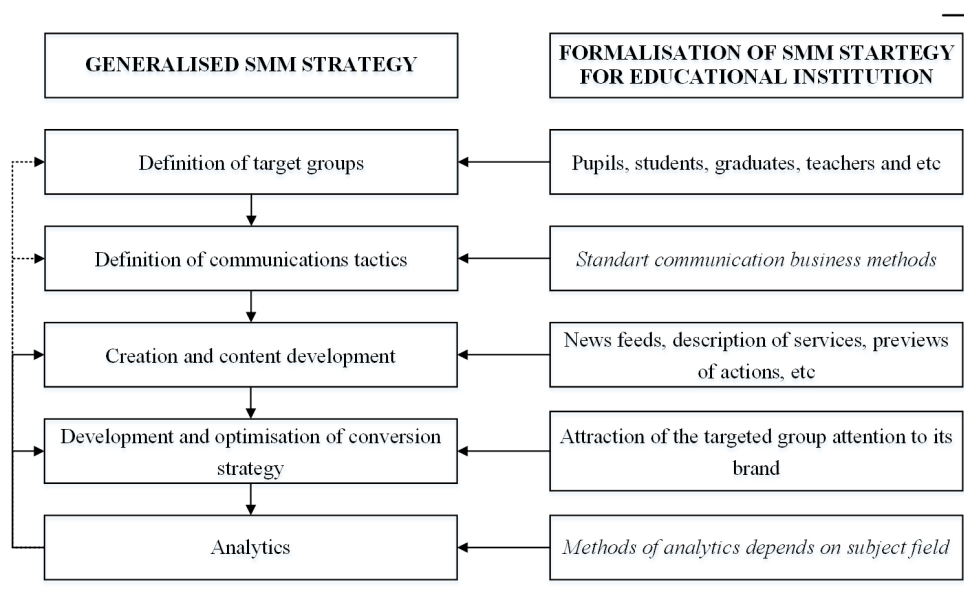
Figure 1: SMM strategy of an educational institution.

and advertising goods and services, making the website more frequently attended.

2   METHODOLOGY

There is a universal concept of SMM strategy. It fully meets the needs of the educational organizations, and some of its stages can be formalized in advance. The SMM strategy of educational institution is presented in Figure 1 (dashed lines show optional links).

Analytics stage is worth noting in particular as it is one of the most important and influencing stages. Monitoring of social networks and blogs, carrying out analytics of brand reference tonality, searching of negative sources on social networks and blogs, carrying out researches on social networks and defining the nature of negative information, analysis of advertising campaign effectiveness, etc. are among primal challenges of social networks analytics.

To have the above tasks solved qualified specialists, efficient tools for monitoring, collecting and analyzing the information obtained from social networks, and also tools for the intellectual analysis of opinions are required.

3   INVESTIGATION

The special place in the field of SMM analytics is allocated to identification of sentiment messages of social networks (positive / negative / neutral). By means of this information it is possible to estimate qualitatively effectiveness of the international activity of educational institution, to reveal a number of the factors influencing the brand image. Up-to-date criteria of sentiment information analysis in the international activity of educational institutions:

- identifications of brand attributes forming negative or positive customer response;
- evaluation of quality and effectiveness of advertising actions;
- effective detection of negative response in social networks;
- evaluation of target audience attitudes to study process of HEI, actions and events carried out by HEI;
- evaluation of target audience attitudes to international activities of HEI.

3.1   Sentiment Analysis

Nowadays the analysis of a tonality is made manually, however, as the information provided in social networks and expressing opinion of authors is in text format, the task of the tonality analysis from the mathematical point of view can be reduced to sentiment analysis task (Gorbushin, 2016). Sentiment analysis is a class of text processing mathematical methods in natural language for identification and analysis of text emotional component. The task of sentiment analysis is a special problem of texts classification and information extraction which lies in the field of the computational (mathematical) linguistics on the edge
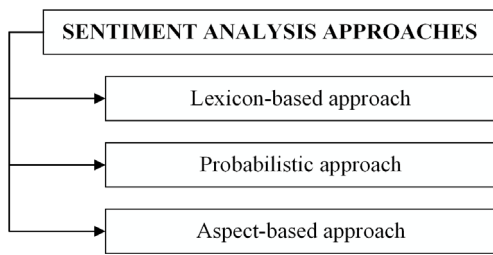
Figure 2: Sentiment analysis approaches.

of linguistics, mathematics, computer science and artificial intelligence (Gorbushin, 2014).

The text processing tool created at the sentiment analysis can be called as linguistic processor. The linguistic processor uses the formal model of language and is language dependent. Approaches to the sentiment analysis task are presented in Figure 2. Let's consider presented approaches to the solution of the sentiment analysis task:

1. Lexicon-based approach (dictionary approach) is based on search of emotive lexicon in text by using the sentiment dictionary (database of polarity words) and application of the boolean rules modeling grammars. It is worth mention that sentiment dictionaries is often used in other approaches. But the main distinctive feature of this approach is that tonality weights or valency of words are considered, on its base total value of text or sentence tonality are calculated.

2. Probabilistic approach is based on the assumption that sequence of words forming tonal unit in one text will have the same tonality in other one. Methods of supervised machine learning,

training of qualifiers on the collections which are in advance marked by experts are the cornerstone of this approach (Pototskiy, 2008). Currently, training datasets for Russian language is not available in free access.

3. Aspect-based approach is based on opinion mining methods. It is the process of extraction and analysis of named entity or aspects of tonality object, which express opinion and can characterize its tonality. Supervised machine learning methods are very popular for solving the problem of aspect extraction. Such methods do not demand large training dataset. However, in a research (Collomb, 2014) aspect approach is considered as a way of overall assessment of text tonality.

Main problems of design and realization of the linguistic processor significantly influencing results of sentiment analysis are presented in Figure 3.

Statistical approach is seen as a most interesting one from the scientific point of view, while aspect-based approach is seen as better option from the qualitative estimation point of view. Linguistic approach does not have any intellectual features due to formalize already accumulated linguistic knowledge. However, rules used in this method can be successfully applied to other approaches to increase classification accuracy. For the analysis of short texts (one sentence) linguistic and aspect approaches are the most efficient methods, as well as some statistical methods. In work (Awadallah, 2012) the authors emphasize that short texts are difficult to be classified due to diverse and rarefying tonality of the linguistic features. At the moment statistical methods are seen as the most effective ones for the large texts analysis (Grinchenkov, 2015).
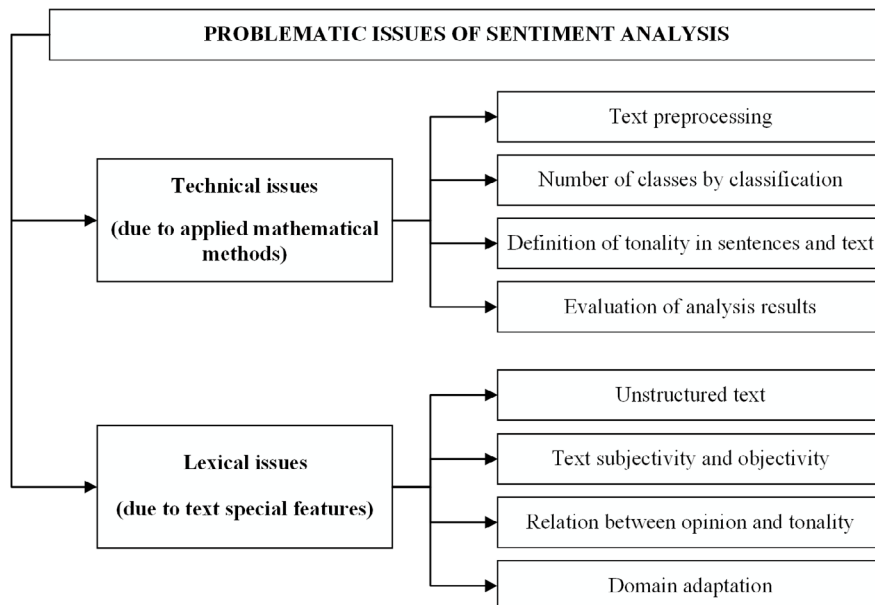

Figure 3: Problematic issues of sentiment analysis.

Problem of statistical methods is development of training dataset with examples from the domain in which the classifier is used. However, methods of linguistic approach have similar problems: sentiment dictionary, compiled for one domain that may not be appropriate for another. The preparatory phase of the statistical approach is the least labor-intensive methods. The main problem of the aspect-based approach is the complexity of drawing up of features extraction algorithms and preparation of ontologies, creating the necessary knowledge about the aspects of objects, which also depends on the subject area.

As for the problem of online marketing of international activities of educational organization it necessary to identify and analyze the opinions of social networks' users. To solve this task, the methods of the aspect approach are seen as the most appropriate.

3.2 Information Model

Taking into account the information above, one can build a model of the analysis of the university international activity management using SMM technology, social network analysis and sentiment analysis. The model is presented in Figure 4.

This model is divided into three main processes: monitoring of public opinion, expressed by the users of social networks; systematic analysis of the opinions and the development and application based on feedback SMM-strategy implemented by previous processes.

3.3 Algorithm

Using the received information, it is possible to develop an algorithm of the software part which is carrying out searching, collecting and the analysis of data from social networks. The algorithm is presented in Figure 5 and contains the following main steps:

- definition of subject domain in social network, keywords;
- search and monitoring of social network publications, including the accompanying data, saving data in the database;
- filtration of the taken data;
- sentiment analysis of the messages stored in the database:
- output the results to the user.

Messages of Vkontakte social network are supposed to be used as an initial data for analysis. To make it possible the keyword database is created for each targeted social network. Afterward the search of users messages are carried out in profile groups using VK API and keywords. As an extra search filters tonal dictionaries could be used. All found messages and extra information ("likes", "reposts") are stored in database.
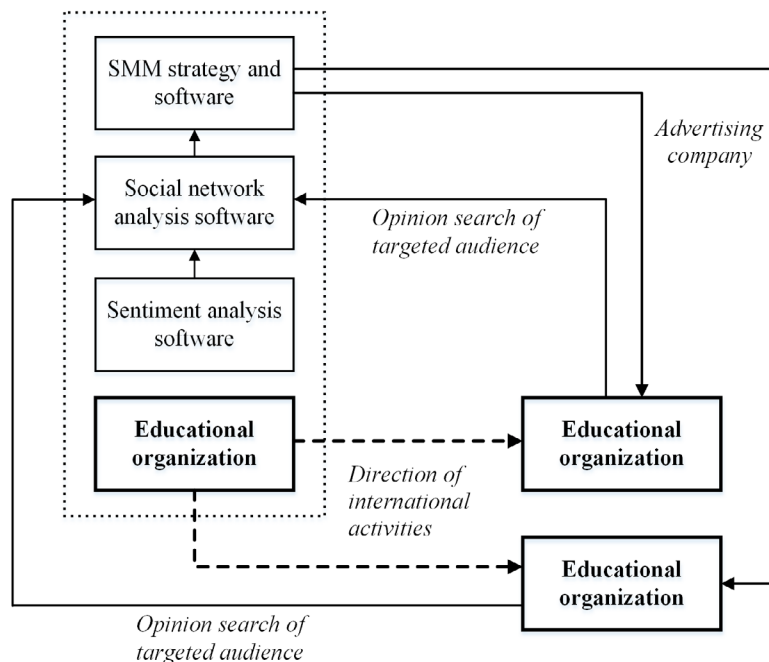

Figure 4: Information model of university international activity management using SMM technology and social network analysis.
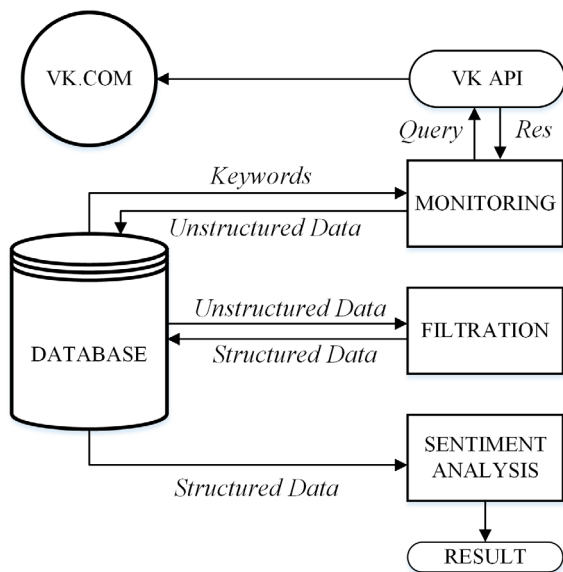
Figure 5: Algorithm.

The findings represent an unstructured text and for the further computer analysis is necessary to make a text preprocessing. At this stage, it could also use tonal dictionaries for filtration of neutral messages.

Next stage is the sentiment analysis of the obtained structured data. The main methods of analyzing of big data are supervised machine learning. They provide a high level of quality and accuracy, simply adapted to the subject area, but require the training dataset.

For the purpose of machine learning algorithms are often used standard vector representations of the text as part of the models «bag of words» or «bag of features». This model can be mathematically described as follows: given multiplicity $\{f_1, \ldots, f_m\}$ of predefined features m, which may appear in a document, $n_i\,(d)$ – the number of iteration of feature f_i in a document d. In this case, each document can be represented as a vector:

$$\bar{d} := (n_1(d), n_2(d), \ldots, n_m(d)) \qquad (1)$$

Consider the basic mathematical methods that use this model.

3.3.1 Support Vector Machines

Support vector machines (support vector networks) is one of the most popular methods of supervised machine learning. The main idea of the method is to convert the original vectors into the space of the higher dimension and to search the separating hyperplane with the maximum interspace in this space. The separating hyperplane is based on two parallel hyperplanes that separate classes.

3.3.2 Naive Bayes Classifier

This classifier is based on the application of Bayes' Theorem with strict (naive) assumptions about independence. For the object of classification are calculated the likelihood function per each of classes, on which are calculated a posteriori class probability. The object is relating to the class where the posteriori probability is maximum (the principle of maximum a posteriori probability).

3.3.3 k-Nearest Neighbors Algorithm

k-Nearest Neighbors algorithm (k-NN) is the simplest metric classifier based on estimation of objects similarity. Classified object belongs to the class, which appertains to the nearest objects of learning sample. k-NN is one of the simplest classification algorithms, therefore, for the real problems it is often ineffective. Beside the accuracy of the classification, the problem of this classifier is the speed of classification: if in the learning sample is N objects, in the test sample is M objects and the dimension of the space is K, then the number of operations O for the classification of the test sample can be evaluated as

$$O = K * M * N \qquad (2)$$

Except to considered methods, there are many other machine learning methods, however, they won't described in detail due to their low popularity.

3.3.4 Comparative Assessment of Methods

In the capacity of the assessment of quality of sentiment analysis is used the standard metrics for the assessment of quality text classification – precision, recall, F-measure (F1). In the research (Vasilyev, 2012) was carried out an experiment to assessment of quality of machine learning methods: SVM – the classifier of support vector machines, GMM – a Bayesian classifier based on mixtures of multivariate normal distributions, ROC – Riccios' classifier, KNN – the classifier k-Nearest Neighbours, VMF – von Mises-Fishers' classifier, TREE – the classifier based on decision trees. Metrics of Recall and Precision were considered separately for positive and negative results. The results of the analysis are presented in Table 1.

Table 1: Comparative assessment of methods of supervised machine learning.

Method	P_p	R_p	P_n	R_n
SVM	0.86	0.99	0.41	0.44
GMM	0.88	0.73	0.27	0.42
ROC	0.92	0.18	0.27	0.8
KNN	0.87	0.78	0.23	0.30
VMF	0.94	0.47	0.31	0.57
TREE	0.90	0.70	0.27	0.30

As you can see from the table above, the quality indicators for negative texts with the use of learning algorithms is noticeably worse than for positive texts. At that, the highest figures demonstrated algorithms: SVM, KNN, TREE.

For the number of classes more than two the results of the support vector method significantly reduced and it turns around mid-table 1. In general, SVM has usually high Precision, but low Recall and it can provide not a very good metric F1. In turn, low Recall is obtained due to the poor recognition of negative lexicon. The above-mentioned researches show that for to improve the quality of the tonal classification by this method it is necessary to apply additional algorithms for searching and processing of negative lexicon.

4 CONCLUSIONS

Application of the presented algorithm allows to obtain the following statistical and analytical information:

- the number of "like", "repost", followers and other quantitative characteristics of social networks and their publications;
- the number of mentions of key words and following word grouping;
- the content analysis of user messages: classification by macro-topics and by emotional tone;
- the clustering of publications according the themes.

The developed model fits into the market model b2c (Business-to-consumer - educational services market) and b2b (Business-to-business - job market). The above information model allows seeing the relationship between these markets and target audience for SMM-strategy, solves the problem of lack of feedback concerning the needs in specialists of a profile, thus solving the acute problem of b2c and b2b markets mismatch, leading to decrease in the efficiency of human resources and reduction in socio-economic development at the regional level.

Information obtained in the analysis can be used not only for the purpose of advertising campaign, but also in the process of optimization and management of educational organization activities (Grinchenkov, 2014).

It is worth noting that the technology of social networks analyses and semantic analysis are the different fast developing fields of research. Current social media monitoring tools are poorly automated and require a lot of human participation, with the analysis being carried out manually. Full-scale intelligent monitoring and views analyzing tools are in the initial stage of development, presenting therefore great scientific and commercial interest.

REFERENCES

Awadallah, R. et al, 2012. PolariCQ: Polarity Classification of Political Quotations. In *Proceedings of the 21st ACM international conference on Information and knowledge management.* ACM, pp. 1945-1949.

Branderburg, A., Federkeli, G., 2007. How to measure the international activities and internationalization of educational institutions. Indicators and key indicators. Release No. 92, Berlin: Center for development of the higher education, 42 p.

Collomb, A., 2014. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. [Online], Available: http://liris.cnrs.fr/Documents/Liris-6508.pdf [05 Dec 2016].

Gorbushin, D.A., Grinchenkov, D.V., Mokhov, V.A., Nguyen, P.H., 2016. System analysis of approaches to solving the problem of identification of the text tonality. In *Proceedings of higher educational institutions. North-Caucasian region. Series: Engineering,* Novocherkassk, no. 2 (196), pp. 36-41.

Gorbushin, D.A., 2014. Analysis of methods of automatic classification of text tonality. In *Proceedings of Scientific-technical conference and exhibition of innovative project*, Novocherkassk, pp. 123-125.

Green, M., 2008. Measurement of internationalization at research universities, New York: American Council of training. 76 p.

Grinchenkov, D.V., Kushchiy, D.N., 2014. The formation of fuzzy collocation on the basis of semantic analysis of the framework curriculum of academic disciplines. In *Proceedings of Problems of modernization of engineering education in Russia: collection of scientific articles on problems of higher school*, Novocherkassk, pp. 298-300.

Grinchenkov, D.V., Kushchiy, D.N., 2015. On selection of keywords in the contents of the framework curriculum of academic disciplines. In *Proceedings of Traditions of the Russian engineering school: yesterday, today, tomorrow: collection of scientific articles on problems of higher school*, Novocherkassk, pp. 114-117.

Pototskiy, S.I., Grinchenkov, D.V., 2008. *Mathematical logic and theory of algorithms for programmers*, URGTU(NPI), Novocherkassk.

Vasilyev V. et. al, 2012. Sentiment classification by fragment rules. In *Kompjuternaja Lingvistika i Intellektual'nye Tehnologii*. Moscow, vol. 11(18), pp. 66–76.