DOI: https://doi.org/10.24113/ijoscience.v7i9.411

Application of Optimization and Machine Learning for Sentiment Analysis

Manitosh Chourasiya
M Tech Scholar
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
chourasiamanitosh@gmail.com

Prof. Devendra Singh Rathore
Assistant Professor
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
devendrarathore2007@yahoo.com

Abstract: **Sentiment analysis is called detecting emotions extracted from text features and is known as one of the most important parts of opinion extraction. Through this process, we can determine if a script is positive, negative or neutral. In this research, sentiment analysis is performed with textual data. A text feeling analyzer combines natural language processing (NLP) and machine learning techniques to assign weighted assessment scores to entities, subjects, subjects, and categories within a sentence or phrase. In expressing mood, the polarity of text reviews could be graded on a negative to positive scale using a learning algorithm.**

The current decade has seen significant developments in artificial intelligence, and the machine learning revolution has changed the entire AI industry. After all, machine learning techniques have become an integral part of any model in today's computing world. However, the ensemble to learning techniques is promise a high level of automation with the extraction of generalized rules for text and sentiment classification activities. This thesis aims to design and implement an optimized functionality matrix using to the ensemble learning for the sentiment classification and its applications.

Keywords: **Lexicon, Aspect level, Sentiment Analysis, Machine Learning.**

I. INTRODUCTION

As is known, the Internet is one of the most used platforms for communication between a person's opinions. This platform explores different areas to express everyone's notes, opinions or emotions. One of the most common means of expressing emotions / opinions is criticism. Whether in the field of cinema, news, products, tweets, etc. In each area, everyone can express themselves through opinions. As most people are online most of the time and express their feelings. Therefore, the database of these assessments grows and thickens every day. According to the latest study, approximately 2.4 billion active online users write and read online worldwide [1].

Although the scientific field as a vast world of journals and conferences is immense, there are over 4000 classified conferences and 5000 classified journals [2]. It should be noted that a large proportion of WWW researchers make their content public and allow researchers, companies, universities and businesses to use and analyze the data. As a result, a large number of studies and researches have observed the trend of online search resources to increase from year to year.

II. LITERATURE REVIEW

Xu et al. [3] proposed an analysis of the sentiment of big data by integrating semantic textual information and neural networks. This approach calculates the weight of individual words and creates a context-sensitive vector. In addition to the feature vector, reference was also made to the data dictionary.

Meyyappan et al. [4] proposed a domain-specific sentiment analysis and called the ConceptNet model. Xu et al. [5] proposed an improved speech feature vector and generated a weighted feature vector to represent mood. The weighted word vectors are then placed into short-term long-term bidirectional memory.

Likewise, C.A. Martin [6] presented a framework in his research that takes short-term networks (LSTM) into consideration as a basis. The ranking of comments was facilitated by this picture and was 89.19% accurate. Even so, the LSTM speed is much slower and less accurate because it takes binary classification into account.

It is not possible to have automatic translations for every language combination. The perception of language in relation to Indian languages has been analyzed by some researchers such as Balamurli and Aditya Joshi [7] in their research analysis. Furthermore, some researchers have proposed a model that offers greater accuracy and reduces the gap by connecting bilingual feelings. Approximately 84% and 72% accuracy was achieved to classify feelings in Marathi and Hindi, respectively.

Bandopadhyaya and Das [8] created approximately 35,805 words and helped develop Senti WordNet, which refers to the Bengali language in the English-Bengali dictionary. To correctly predict the word feelings, four strategies have been suggested in their work. The very first strategy was an interactive game which, with its popularity, translated into words. To predict polarity with the second strategy, bilingual dictionaries in other Indian languages and English were used. The third strategy involved predicting polarity through antonym-synonym and WordNet relationships. Predrogated corpora were used in the third strategy to discover polarities.

III. METHODOLOGY

The proposed methodology are the analyzing of the sentiment of textual data. The proposed global algorithm is illustrated in fig. 1, which comprises four steps, Extraction of data from datasets, preprocessing, feature selection and the classification.

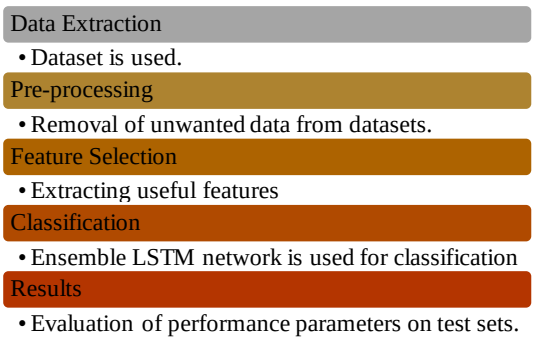

Fig. 1 Proposed Flow Diagram

A. *Dataset Extraction*

The research methodology is a dataset is the created. For the data preparation, data are first of all collected data sets. Three datasets are used in this work, i. H. IMDB, Semeval and SST.

B. *Data Pre-processing*

It is quite necessary to clean up the raw data collected from various resources. This step is known as data preprocessing. Unnecessary terms in data records such as commas and special characters are removed during preprocessing because they do not contribute to sentiment values in either the sentence or the document. The collected dataset is parsed entities for entities and unnecessary entities such as URLs, special characters, commas, etc. are cleared and a clean dataset is prepared for further processing.

C. *Feature Selection*

Figure 2 shows the feature selection process for the proposed methodology, which is explained below:

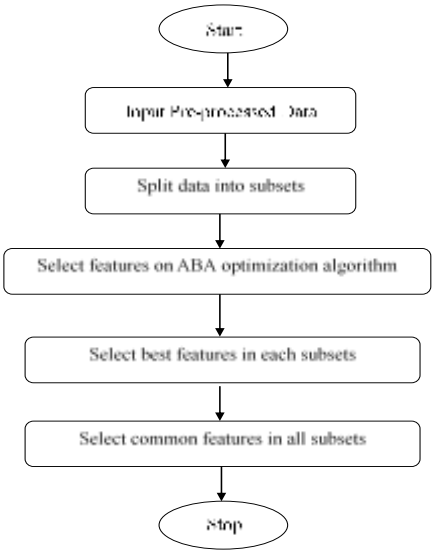

Fig. 2 Flowchart of feature Selection

To generate a vector of characteristics related to a positive evaluation and a negative evaluation of the data, we refer to the dictionary of sentiment data. In this paper, the sentiment score is calculated by adding the positive and negative scores of each word in the entire sentence.

Inspired by the strategy of finding identified forest partners, the optimization of artificial butterflies was developed. Spotted forests prefer to live on the edge of the forest, where the sun shines on the trees and forms many sunspots. The butterfly population is sorted and divided into two groups based on their physical form. The most suitable butterflies are sunspot butterflies and the rest are canopy butterflies, and a different flight strategy is used for each group.

Two modes compose the ABO algorithm:

Sunspot mode

Canopy mode

Some rules for butterflies in the ABO algorithm are given as follows:

- To increase the likelihood of encountering female butterflies, all male butterflies try to fly to a better place called a sunspot.
- To occupy a better sunspot, each sunspot butterfly always tries to fly to the neighboring sunspot.
- Each butterfly in the canopy constantly flies to each sunspot butterfly to fight for the sunspot.

Let the P = {p_1 , p_2 , . . . , p_m} = Population of the butterflies.

The following strategy is used for sunspot mode or sail mode. Each butterfly flies to a randomly selected butterfly as follows:

$$P_i^{n+1} = (P_i^n - P_k^n)\beta \qquad (3.1)$$

Where, i= ith butterfly
n =iteration
β =random generated number between [1, −1]
k =randomly selected butterfly (k≠i)

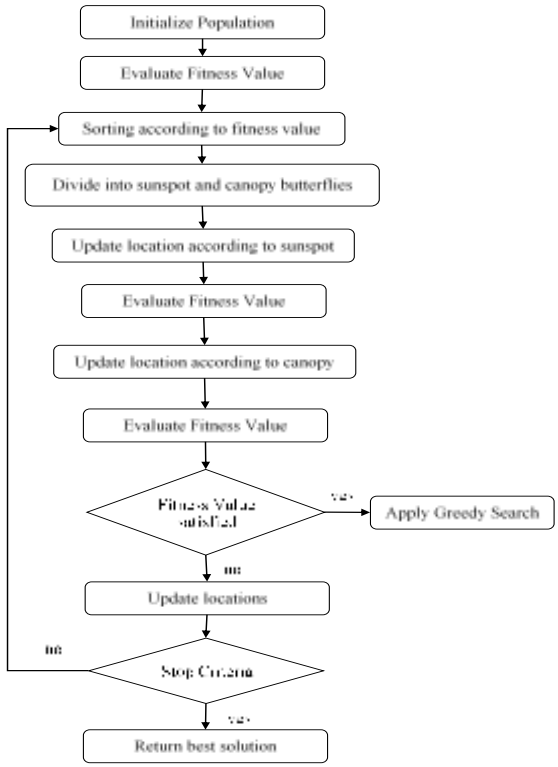

Fig. 3 Artificial Butterfly Optimization Algorithm

Each butterfly flies to a randomly selected solar butterfly as follows:

$$P_i^{n+1} = P_i^n + \frac{P_k^n - P_i^n}{x_k^n - P_i^n})(Ub - Lb)s\beta \qquad (3.2)$$

Where, Ub= upper of bound

Lb= Lowe bound

The s parameter of decreases linearly in from 1 to s_e , as follows:

$$s = 1 = (1 - S_e)\frac{n}{N} \qquad (3.3)$$

where N = Max iteration

D. Classification

Classification algorithms are used to classify data values into different categories. A set of LSTMs are used in this work and the review data is categorized into different polarities of opinion. In this section, the dataset is converted to a vector of words before classification and placed in the classifier. The extracted word vectors are inserted into the model as an initial value. Lexicon features and deep aspect level features are extracted and train the classifier for the polarity decision.

In this exploration work, the informational collection is handled and the reproduction is performed with the proposed calculation. Set classifiers are used for performance evaluation and are discussed below:

1 Ensemble LSTM Training

Once the data samples have been grouped, they are inserted into a network of deep classifiers (Ensemble LSTM). As in the traditional LSTM network, the last level is the Softmax level, which is less efficient. It is then replaced by a random forest classifier. This helps in analyzing the type of log data. 4 represents the block diagram of the proposed training architecture. Long-term memory (LSTM) is one of the types of recurrent neural networks, a deep learning technique. Basically, each node in the LSTM network consists of four units, called the memory cell, input unit, output unit, and forgotten unit. The function of the memory unit is to store internal parameter values for a certain time and the other three units are control units that control the flow of data values to evaluate the output. At each instant of the sample t an input port it, an forgetful port ft, an output port ot and a memory cell Ct are initialized.

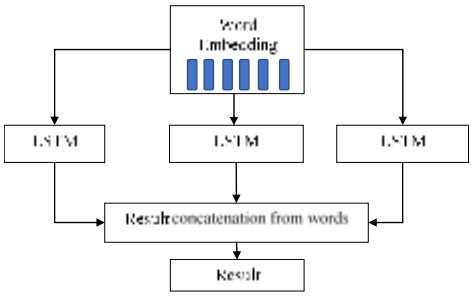

Fig. 4 Feature Classification for Proposed Work

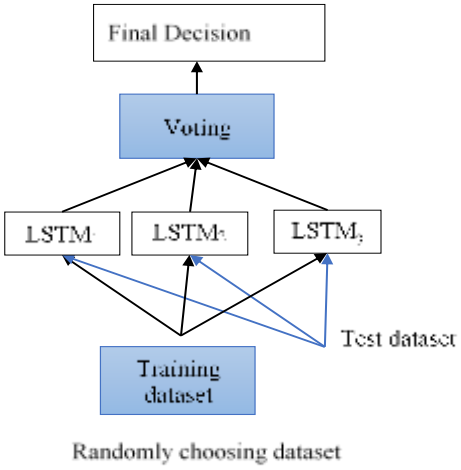

Fig. 5 Proposed Training Architecture

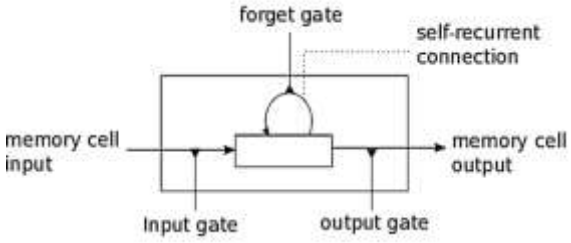

Fig. 6 LSTM Units

All of these are used to calculate the hidden output layer ht as follows:

$$f_t = \sigma(W_f * x_t + U_f * h_{t-1} + b_f) \quad (3.3)$$

$$i_t = \sigma(W_i * x_t + U_i * h_{t-1} + b_i) \quad (3.4)$$

$$\hat{C}_t = tanh(W_c * x_t + U_c * h_{t-1} + b_c) \quad (3.5)$$

$$C_t = i_t * \hat{C}_t + f_t * C_{t-1}) \quad (3.6)$$

$$o_t = \sigma(W_o * x_t + U_o * h_{t-1} + b_o) \quad (3.7)$$

$$h_t = o_t * \tanh(C_t)) \quad (3.8)$$

IV. RESULT

This part of the scientific and exploratory portrayal of the proposed procedure for opinion investigation. The reproduction is performed utilizing the MATLAB stage to assess execution. For the simulation results, the work focuses on extracting lexical functionality and aspect-level functionality for sentiment analysis from scores. To run the simulation, reviews are pulled from various domains and collected from datasets such as movie review data and Stanford Sentiment Treebank review data. While some reviews were collected from the Tripadvisor website.

A. Performance Parameters

Accuracy

This is one of the most important parameters in determining the efficiency of the classifier. Represents all correctly classified outputs, positive or negative. The mathematical representation of precision is like in the equation. (1):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Where,

TP = True Positive (represents the total number of test samples that should be positive and whose actual labeling is also positive).

TN = True Negative (represents the total number of test samples expected to be negative and whose actual labeling is negative).

PF = False Positive (represents the total number of test samples that were predicted to be positive and were actually marked negative).

FN = False Negative (represents the total number of test samples that are expected to be negative and are in fact marked as positive).

Precision

Likewise, another parameter for performance evaluation is precision, which determines all correct positive classifications of all expected positive samples. Mathematically, accuracy is like in the equation clock (2):

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

Recall

Another performance evaluation parameter that determines positive prediction out of all actual positive samples is termed as recall. Mathematically it is represented as in eqn (4):

$$Recall = \frac{TP}{(TP + FN)} \qquad (3)$$

F_Measure

The harmonic mean between recall and precision is called f_measure. Mathematically, it becomes like in the equation clock (4):

$$F_{measure} = \frac{2 * Recall * Precision}{(Recall + Precision)} \qquad (4)$$

B. Result Analysis

Table 1 shows the presentation assessment of the proposed calculation with and without informational collection enhancement. From the examination of the outcomes, it was investigated that the grouping with the advancement accomplished the best outcome.

Table 1 Performance Evaluation of Proposed Algorithm

Algorithms	Accuracy
With Optimization	93.45
Without Optimization	85.34

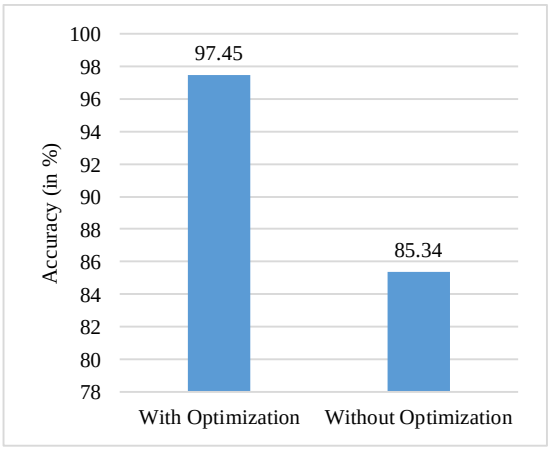

Fig. 7. Performance Comparison of Accuracy with and Without Optimization

The performance evaluation of the methodology proposed in this thesis is presented in Table 2. Because the sentiment analysis in this thesis is evaluated using machine learning. Performance is rated for accuracy, precision, recall and measurement f. Various test kits are used to demonstrate the effectiveness of the model.

Table 2: Performance Evaluation on IMDB Data from Different Sources

	IMDB	Semeval	SST	Average
Accuracy	91.62	91.29	92.85	91.92
Recall	92.65	93.23	92.89	92.923
Precision	91.34	93.20	93	91.84
F_Measure	92.49	91.76	93.97	92.74

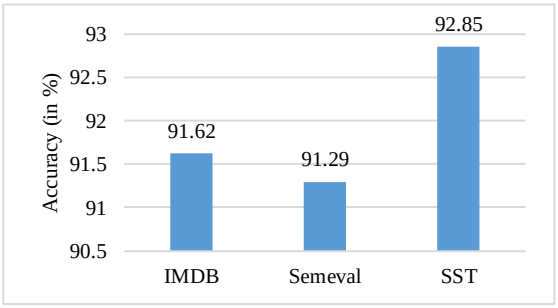

Fig. 8 Performance Evaluation of Accuracy

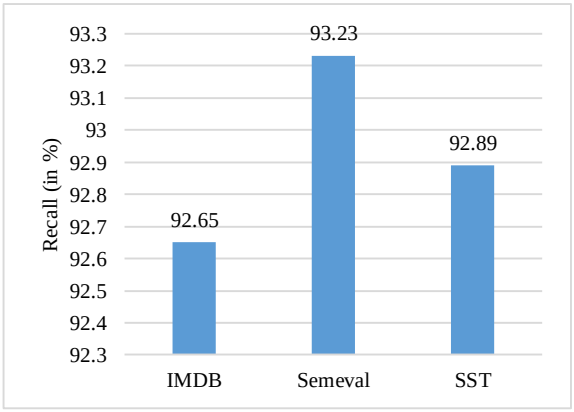

Fig. 9 Performance Evaluation of Recall

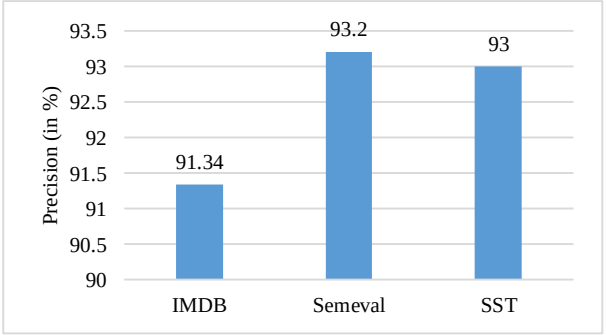

Fig. 10 Performance Evaluation of Precision

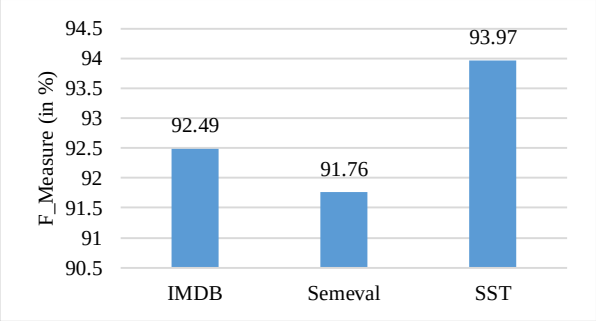

Fig. 11 Performance Evaluation of Accuracy F_Measure

Table 3 and table 4 represents the comparative performance evaluation with research work presented by Lin et al. (2020). Lin et al. (2020) presented bidirectional LSTM network for sentiment analysis on different domain.

Table 3: Comparative Accuracy Evaluation with Existing Work

Datasets	Lin et al. (2020)	Proposed
IMDB	89.5	91.62
Semeval	80.4	91.29
SST	73.5	92.85
Average	81.13	91.92

Table 4 Comparative Precision Evaluation with Existing Work

Datasets	Lin et al. (2020)	Proposed
IMDB	92.9	91.34
Semeval	70.3	93.20
SST	73.2	93
Average	78.8	92.51

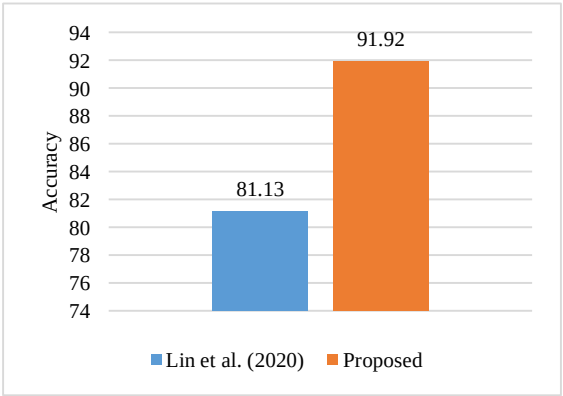

Fig. 12 Accuracy Comparative Performance Evaluation

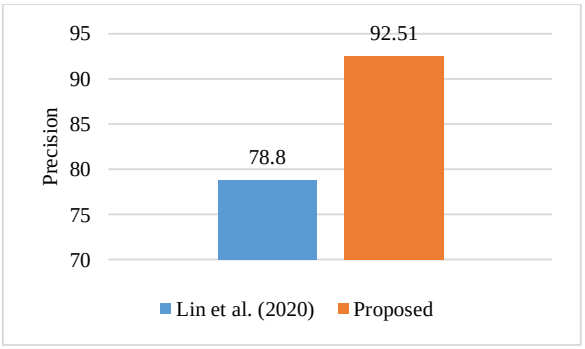

Fig. 13 Precision Comparative Performance Evaluation

V. CONCLUSION

Of all the means of communication, the textual one is today one of the most privileged media. People express their feelings by sharing their opinion. In most apps, people express their feelings by sharing reviews, and people trust each app by analyzing reviews, be it products, movies, apps, news, etc. it is determined by the analysis of the reviews. Most of the research these days is aimed at analyzing these criticisms and determining their polarity towards positive and negative perceptions. This analysis is classified under natural language processing. Additionally, machine learning contributed the most to this analysis.

In recent years, the rise of opinion content on the Internet has broadened the reach of data analytics and created new opportunities for new challenges. The analysis of user-generated activity is very important and time-consuming and must be targeted. The phrases extracted from the text elements would help determine the polarity of the review as positive or negative. In this thesis, a sentiment analysis framework optimized for different test datasets is proposed. In this thesis, a feature reduction technique is proposed to solve the scalability problem that arises with the growth of the feature set, which uses swarm intelligence, the so-called artificial butterfly algorithm and a complete machine approach with an Optimized selection of features contains. Analysis of the results shows improvements over existing work.

VI. FUTURE WORK

In future work, this research will focus on integration for feature correlation analysis in order to select optimal features. In future work, this work will be expanded to include news article reviews, review analysis, and tips to improve the functionality of future work.

REFERENCES

[1] Y. Lin, J.Li, L. Yang, and H. Lin. “Sentiment Analysis with Comparison Enhanced Deep Neural Network”. IEEE Access.8, pp. 78378-78384, 2020.

[2] L. Yang, Y.Li, and R. S. Sherratt. “Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning”. IEEEAccess. vol8, pp. 23522-23530, 2020.

[3] G. Xu, Z. Yu, Z. Chen, and H.Yao, “Sensitive Information Topics-Based Sentiment Analysis Method for Big Data”. IEEE Access. vol7, pp.96177-96190, 2019.

[4] V. Ramanathan and T.Meyyappan “Twitter Text Mining for Sentiment Analysis on People’s Feedback about Oman Tourism. International Conference on Big Data and Smart City (ICBDSC). pp. 1-5, 2019.

[5] G.Xu, Y.Meng, X. Qiu, and X. Wu.. “Sentiment Analysis of Comment Texts Based on BiLSTM”. IEEEAccess. vol.7, pp. 51522-51532,2019.

[6] C. A. Martín, J. M. Torres, and S. Diaz. “Using Deep Learning to Predict Sentiments: Case Study in Tourism”. Hindawi Complexity, 2018.

[7] Balamurali A R, Aditya Joshi, Pushpak Bhattacharyya. (2012). Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets. Coling. pp. 73-78,2012.

[8] Amitava Das, Sivaji Bandopadaya.. SentiWordnet for Bangla. Knowledge Sharing Event -4. 2,2010.

[9] Annett, Michelle, and GrzegorzKondrak.. “A comparison of sentiment analysis techniques: Polarizing movie blogs”. Advances in artificial intelligence. Springer Berlin Heidelberg. pp. 25-35,2010.

[10] Bai, Xue. “Predicting consumer sentiments from online text”. Decision Support Systems”. 50(4). pp. 732-742,2011.