ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Text Document Classification by using WordNet Ontology and Neural Network

#1Ms. Tejashri Mane, #2Ms. Dhanashree Ghone, #3Mr.Prasad Khade, #4Ms. Manisha Gawade, #5Prof.Muzaffar A.R. Shabad

Department of Computer Engineering, Savitribai Phule Pune University, Sinhgad Academy of Engineering, Kondhwa, Pune, India #1234

ABSTRACT: every day the mass of information available, merely finding the relevant information is not the only task of automatic text classification systems. The main problem is to classify which documents are relevant and which are irrelevant. The Automated text classification consists of automatically organizing data. We propose a method of automatic text classification using Convolutional Neural Network based on the disambiguation of the meaning of the word we use the WordNet ontology and word embedding algorithm to eliminate the ambiguity of words so that each word is replaced by its meaning in suitable context. The closest ancestors of the senses of all the words in a given document are selected as folders for the specified document.

KEYWORDS: neural network, classification, feature selection, model selection, WordNet.

I. INTRODUCTION

Every day the mass of information available to us increases. This information would be irrelevant if our ability to productively get to did not increment too. For most extreme advantage, there is need of devices that permit look, sort, list, store and investigate the accessible information. One of the promising region is the automatic text categorization. Envision ourselves within the sight of impressive number of texts, which are all the more effectively available on the off chance that they are composed into classes as per their topic. Obviously one could request that human read the text and arrange them physically. This assignments is hard if done on hundreds, even a huge number of texts. Thus, it appears to be important to have a computerized application, so here automatic text categorization is presented. An increasing number of data mining applications involve the analysis of complex and structured types of data and require the use of expressive pattern languages. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the neural network.

Unfortunately, existing "upgrading" approaches, especially those using Logic Programming techniques, often suffer not only from poor scalability when dealing with complex database schemas but also from unsatisfactory predictive performance while handling noisy or numeric values in real-world applications. However, "flattening" strategies tend to require considerable time and effort for the data transformation, result in losing the compact representations of the normalized databases, and produce an extremely large table with huge number of additional attributes and numerous NULL values (missing values). As a result, these difficulties have prevented a wider application of multi relational mining, and post an urgent challenge to the data mining community. To address the above mentioned problems, this article introduces a classification approach where neither "upgrading" nor "flattening" is required to bridge the gap between propositional learning algorithms and relational.

In Proposed approach, Data analysis techniques, such as classification it can be used to identify subsets of data instances with common characteristics. Users can explore the data by examining some instances in each group instead of rather than examining the instances of the complete data set. This allows users to focus efficiently on large relevant subsets Data sets, in particular for document collections. In particular, the descriptive grouping consists of automatic

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

grouping sets of similar instances in clusters and automatically generate a description or a synthesis that can be interpreted by man for each group. The description of each cluster allows a user determine the relevance of the group without having to examine its content For text documents, a description suitable for each group can be a multi-word tag, an extracted title or a list of characteristic words . The quality of the grouping it is important, so that it is aligned with the idea of likeness of the user, but it is equally important to provide a user with a brief and informative summary that accurately reflects the contents of the cluster

II. RELATED WORK

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers.

In this section, we briefly review the related work on Text classification and their different techniques.

J.-T. Chien, describe the "Hierarchical theme and topic modeling," in that Taking into account hierarchical data sets in the body of text, such as words, phrases and documents, we perform structural learning and we deduce latent themes and themes for sentences and words from a collection of documents, respectively. The relationship between arguments and arguments in different data groupings is explored through an unsupervised procedure without limiting the number of clusters. A tree branching process is presented to draw the proportions of the topic for different phrases. They build a hierarchical theme and a thematic model, which flexibly represents heterogeneous documents using non-parametric Bayesian parameters. The thematic phrases and the thematic words are extracted. In the experiments, the proposed method is evaluated as effective for the construction of a semantic tree structure for the corresponding sentences and words. The superiority of the use of the tree model for the selection of expressive phrases for the summary of documents is illustrated [1].

Bernardini, C. Carpineto, and M. D'Amico, describe the "Full-subtopic retrieval with keyphrase-based search results classification," in that Consider the problem of restoring multiple documents that are relevant to the individual sub-topics of a given Web query, called "full child retrieval". To solve this problem, they present a new algorithm for grouping search results that generates clusters labelled with key phrases. The key phrases are extracted generalized suffix tree created by the search results and merge through a hierarchical agglomeration procedure improved grouping. They also introduce a new measure to evaluate the performance of full recovery sub-themes, namely "look for secondary arguments length under the sufficiency of k documents". they have used a test collection specifically designed to evaluate the recovery of the sub-themes, they have found that our algorithm has passed both other classification algorithms of existing research results as a method of redirecting search results underline the diversity of results (at least for k> 1, that is when they are interested in recovering more than one relevant document by sub-theme) [2].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, describe the "Self-organization of a massive document collection," this paper describes the implementation of a system that can organize large collections of documents based on textual similarities. It is based on the self-organized map (SOM) algorithm. Like the feature vectors for documents, the statistical representations of their vocabularies are used. The main objective of our work was to resize the SOM algorithm in order to handle large amounts of high-dimensional data. In a practical experiment, they mapped 6 840 568 patent abstracts in a SOM of 1.002.240 nodes. As characteristic vectors, we use vectors of 500 stochastic figures obtained as random projections of histograms of weighted words [3].

K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, describe the "A hierarchical monothetic document classification algorithm for summarization and browsing search results," in that Organizing Web search results in a hierarchy of topics and secondary topics makes it easy to explore the collection and position the results of interest. In this paper, they propose a new hierarchical monarchic grouping algorithm to construct a hierarchy of topics for a collection of search results retrieved in response to a query. At all levels of the hierarchy, the new algorithm progressively identifies problems in order to maximize coverage and maintain the distinctiveness of the topics. They

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

refer to the algorithm proposed as DisCover. The evaluation of the quality of a hierarchy of subjects is not a trivial task, the last test is the user's judgment. They have used various objective measures, such as coverage and application time for an empirical comparison of the proposed algorithm with two other monotetic grouping algorithms to demonstrate its superiority. Although our algorithm is a bit more computationally than one of the algorithms, it generates better hierarchies. Our user studies also show that the proposed algorithm is superior to other algorithms as a tool for summary and navigation [4].

Ying Liu1, Peter Scheuermann2, Xingsen Li1, and Xingquan Zhu, describe the “Using WordNet to Disambiguate Word Senses for Text Classification,” in that they propose an automatic method of text classification. Based on the disambiguation of the meaning of words. We use the "bell" algorithm to eliminate the word ambiguity so that every word is replaced by its meaning in context. The closest ancestors of the senses of all words without stopping in a given document Selected as classes for the specified document [5].

S. Dumais, J. Platt, D. Heckerman, and M. Sahami, describe the “Inductive learning algorithms and representations for text categorization,” in that Text categorization the assignment of natural language texts to one or more predefined categories based on their content is an important component in many information organization and management tasks. They compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real-time classification speed and classification accuracy. They also examine training set size, and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. Linear Support Vector Machines (SVM) are particularly promising because they are very accurate, quick to train and quick to evaluate [6].

R. Kohavi and G. H. John, describe the “Wrappers for feature subset selection, “In that the feature subset selection problem, a learning algorithm is faced with the problem of selecting a relevant subset of features upon which to focus its attention, while ignoring the rest. To achieve the best possible performance with a particular learning algorithm on a particular training set, a feature subset selection method should consider how the algorithm and the training set interact. They explore the relation between optimal feature subset selection and relevance. Our wrapper method searches for an optimal feature subset tailored to a particular algorithm and a domain. They study the strengths and weaknesses of the wrapper approach and show a series of improved designs. They compare the wrapper approach to induction without feature subset selection and to Relief, a filter approach to feature subset selection. Significant improvement in accuracy is achieved for some datasets for the two families of induction algorithms used: decision trees and Naive-Bayes [7].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero,and A. Saarela, describe the “Self-organization of a massive document collection,” This paper describes the implementation of a system that is able to organize vast document collections according to textual similarities. It is based on the self-organizing map (SOM) algorithm. As the feature vectors for the documents statistical representations of their vocabularies are used. The main goal in our work has been to scale up the SOM algorithm to be able to deal with large amounts of high-dimensional data. In a practical experiment we mapped 6 840 568 patent abstracts onto a 1 002 240-node SOM. As the feature vectors we used 500-dimensional vectors of stochastic figures obtained as random projections of weighted word histograms [8].

Q. Mei, X. Shen, and C. Zhai, describe the “Automatic labeling of multinomial topic models,” In this paper, they propose probabilistic approaches to automatically labeling multinomial topic models in an objective way. They cast this labeling problem as an optimization problem involving minimizing Kullback-Leibler divergence between word distributions and maximizing mutual information between a label and a topic model. Experiments with user study have been done on two text data sets with different genres. The results show that the proposed labeling methods are quite effective to generate labels that are meaningful and useful for interpreting the discovered topic models. Our methods are general and can be applied to labeling topics learned through all kinds of topic models such as PLSA, LDA, and their variations [9].

K. Lagus and S. Kaski, describe the “Keyword selection method for characterizing text document maps,” in that Characterization of subsets of data is a recurring problem in data mining. They propose a keyword selection method

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

that can be used for obtaining characterizations of clusters of data whenever textual descriptions can be associated, with the data. Several methods that cluster data sets or form projections of data provide an order or distance measure of the clusters. If such an ordering of the clusters exists or can be deduced, the method utilizes the order to improve the characterizations. The proposed method may be applied, for example, to characterizing graphical displays of collections of data ordered e.g. with the SOM algorithm. The method is validated using a collection of 10,000 scientific abstracts from the INSPEC database organized on a WEBSOM document map [10].

III. PROPOSED APPROACHES

In Proposed System training is creation of train data set using which classification of unknown data in predefined categories is done. Here a learning system is created using neural network approach. It is a supervised learning where unlabeled data (test data) is classified using labelled data (training dataset). Training data is always a labelled dataset based on its features.

Project had considered no of scientific papers form different publication of different domains for creating training dataset. These papers are input for creating training dataset. This input is first preprocessed and most informative features are extracted using TF/IDF and Word embedding word sense algorithm. Ten different domains from market are identified and then extracted feature and have to put to corresponding domain where each domain is considered as one class that which is used for labeling test dataset in testing part and features are considered as nodes. Once training part is completed, all features of respective domains are get updated in corresponding tables in database.

A. System Architecture

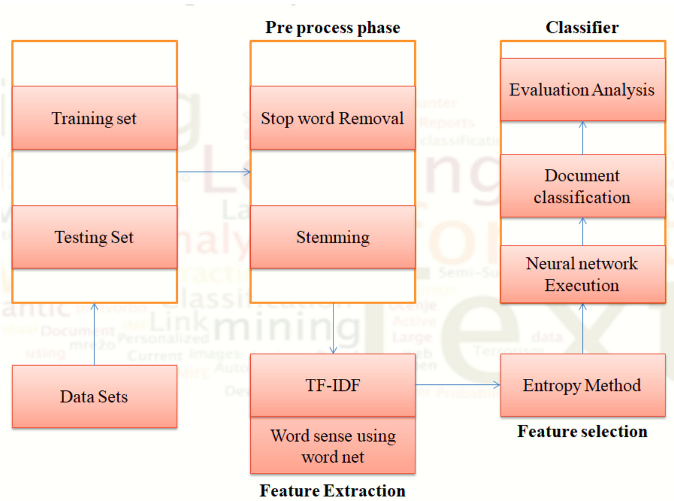


FIG 1: SYSTEM ARCHITECTURE

B. Methodology

Step1: Remove all stop words such as the, a, an etc., and also all functional words such as adverbs, preposition, conjunction etc, from the text of all the documents in the given set.

Step 2: Necessary morphological analysis to extract the root words from the given set of texts is carried and remove all the repetition of the root words.

Step 3: We define an elimination factor TF/IDF for each word as = Number of occurrence in its own context / Total number of occurrences in all contexts.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Step 4: Select the features to be all the remaining words of all the documents in the given set of text documents.

Step 5: Create one pattern vector for each document with the features (i.e. words) selected in step 3. The numeric value for each component of such vector would be the number of occurrence for the particular word corresponding to that component in the given document. Note: Step 3 eliminates all those words that are used al- most to the similar extent in all the given classes of documents. Thus these words have almost no discriminatory significance so far classification is concerned.
Convolutional neural network (CNN) is a kind of typical artificial neural network.In this kind of network, the output of each layer is used as the input of the nextlayer of neuron. Multi-layer convolution operation is used to transform the resultsof each layer by nonlinear until the output layer. In general, the convolution neural network model used in text analysis.

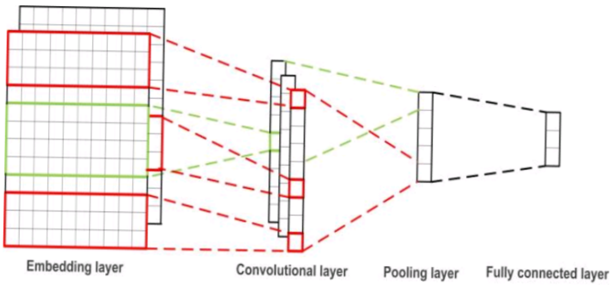


fig 2. convolutional neural network

Which includes four parts: embedding layer,convolutional layer, pooling layer and fully connected layer. Compared with thetraditional models for image analysis, the difference is that the input layer of theCNN model used in text analysis is the word vector.WordNet is a manually-constructed lexical system developed by George Miller atthe Cognitive Science Laboratory at Princeton University. Itreacts how humanbeings organize their lexical memories. The basic building block of WordNet issynset consisting of all the words that express a given concept. Synsets, which senses are manually classified into, denote synonym sets. Within each synset, the senses, although from different keywords, denote the same meaning.

IV. RESULTS AND DISCUSSION

In experimental results, we evaluate the proposed system on student conference papers datasets this available on internet. We compare the accuracy of existing system results with proposed system.

The experimental result evaluation, we have notation as follows:
TP: True positive (correctly predicted number of instance)
FP: False positive (incorrectly predicted number of instance),
TN: True negative (correctly predicted the number of instances as not required)
FN false negative (incorrectly predicted the number of instances as not required),
On the basis of this parameter, we can calculate four measurements
Accuracy = TP+TN÷TP+FP+TN+FN
Precision = TP ÷TP+FP
Recall= TP÷TP+FN
F1-Measure = 2×Precision×Recall ÷Precision+ Recall.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

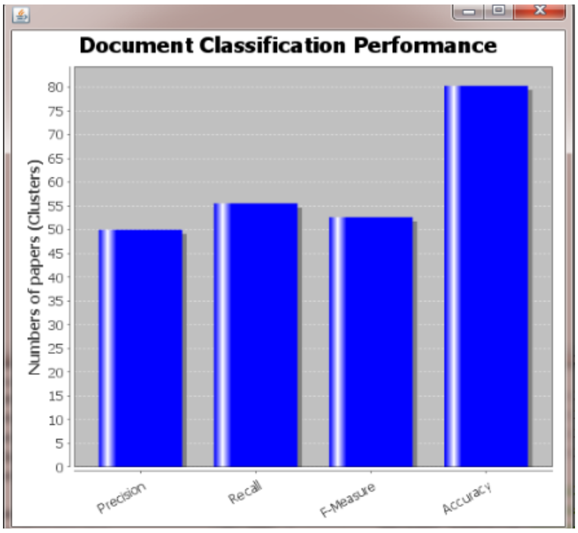


Fig 3. Analysis Graph

Parameters	Percentage
Precision	48.6
Recall	57.1
F-Measure	51.8
Accuracy	78.4

Table 1: accuracy analysis

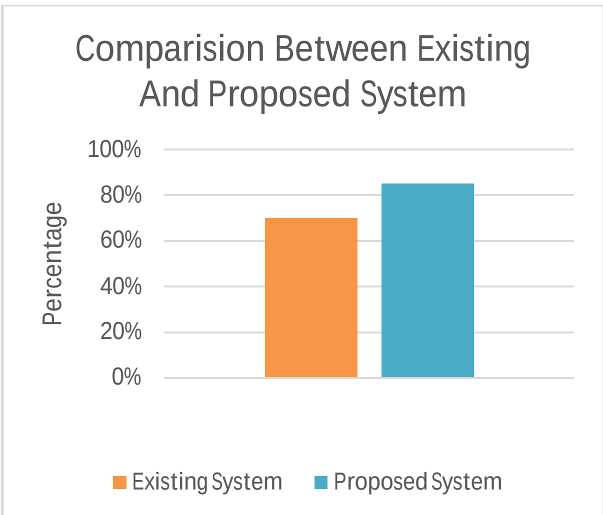


Fig 4. Graph

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

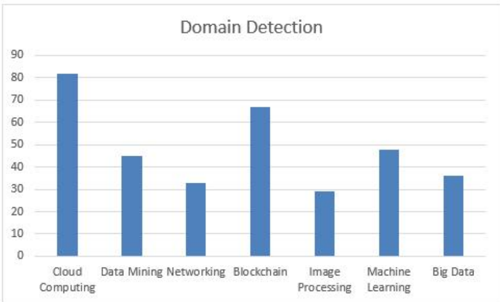
Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

A. *Comparison Table:*

Sr.No	Existing Result	Proposed Result
1	65 to 70%	78.4%

Table 2.comparative result



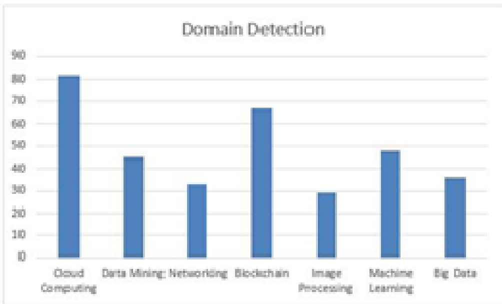


Fig. 5. Accuracy Graph

Sr. No.	Domain Name	Paper count
1	Cloud Computing	82
2	Data Mining	45
3	Networking	34
4	Blockchain	65
5	Image Processing	29
6	Machine Learning	46
7	Big Data	36

Table 3:Domain Detection Result

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

V. CONCLUSION

Proposed Text classification as two coupled predictions activity choose a grouping that is predictive of features. Use predictive performance as a goal criterion, classification parameters the number of function: they are chosen from the model selection. With the result solution, each group is described by a minimum subset of features necessary to predict if an instance belongs to the data our hypothesis is that even a user will be able to predict membership in the group of documents using the features selected by the WordNet and neural network. Given Some relevant requirements, a user can quickly identify that probably contain relevant documents

REFERENCES

[1] J.-T. Chien, “Hierarchical theme and topic modeling,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 565–578, 2016.

[2] Bernardini, C. Carpineto, and M. D’Amico, “Full-subtopic retrieval with keyphrase-based search results clustering,” in IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intelligent Agent Technol., 2009, pp. 206–213.

[3] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, “Self-organization of a massive document collection,”IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.

[4] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, “A hierarchical monothetic document clustering algorithm for summarization and browsing search results,” in Proc. Int. Conf. World Wide Web, 2004, pp. 658–665.

[5] Ying Liu1, Peter Scheuermann2, Xingsen Li1, and Xingquan Zhu1 Using WordNet to Disambiguate Word Senses for Text Classification.

[6] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representations for text categorization,” in Proc. Int. Conf. Inform. Knowl. Manag., 1998, pp. 148–155.

[7] R. Kohavi and G. H. John, “Wrappers for feature subset selection,”Artif. Intell., vol. 97, no. 1, pp. 273–324, 1997.

[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero,and A. Saarela, “Self-organization of a massive document collection,”IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.

[9] Q. Mei, X. Shen, and C. Zhai, “Automatic labeling of multinomial topic models,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2007, pp. 490–499.

[10] [10]K. Lagus and S. Kaski, “Keyword selection method for characterizing text document maps,” in Int. Conf. Artificial Neural Networks (ICANN), 1999, pp. 371–376.

[11] Kewei Shen; Xian Chen."A blended feature selection method in text classification", International Conference on Cyberspace Technology (CCT 2013). 2013: 573 - 576. [12]Xiaofei Zhou, Li Guo. "Latent Factor SVM for Text Categorization". IEEE International Conference on Data Mining Workshop (ICDMW). 2014捖Pages: 105 - 110.

[13]Zhe Gao; Yajing Xu,et al. "Improved information gain-based feature selection for text categorization". International Conference on Wireless Communications,Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE). 2014: 1 - 5.

[14]Dai Li, Yi L."Automatic text categorization using a system of highprecision and high-recall models". Dai Li; Murphey, Y.L.捖 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). 2014: 373 - 380.

[15]Nouri, V.; Akbarzadeh-T,et al. "A hybrid type-2 fuzzy clustering technique for input data preprocessing of classification algorithms". 2014 IEEE International Conference on Fuzzy Systems (FUZZIEEE). Kunming: IEEE Computer Society. 2014: 1131 - 1138.

[16]Nouri, V.; Akbarzadeh-T."A hybrid type-2 fuzzy clustering technique for input data preprocessing of classification algorithms": 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2014: 1131 - 1138.