2018-11

Opinion Mining on Small and Noisy Samples of Health-Related Texts

John Cardiff
Technological University Dublin, john.cardiff@tudublin.ie

Liliya Akhtyamova
Technological University Dublin

Mikhail Alexandrov
Autonomous University of Barcelona

See next page for additional authors

Recommended Citation

Akhtyamova L., Alexandrov M., Cardiff J., Koshulko O. (2019) .Opinion mining on small and noisy samples of health-related texts. In (Shakhovska N., Medykovskyy M.) (Eds). *Proceedings of Advances in Intelligent Systems and Computing III. CSIT 2018. Advances in Intelligent Systems and Computing* vol 871. Springer

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Advances in Intelligent Systems and Computing III	
Series Title		
Chapter Title	Opinion Mining on Small and Noisy Samples of Health-Related Texts	
Copyright Year	2019	
Copyright HolderName	Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Akhtyamova**
	Particle	
	Given Name	**Liliya**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Institute of Technology Tallaght
	Address	Dublin, Ireland
	Email	liliya.akhtyamova@postgrad.ittdublin.ie
Author	Family Name	**Alexandrov**
	Particle	
	Given Name	**Mikhail**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Autonomous University of Barcelona
	Address	Barcelona, Spain
	Division	
	Organization	Russian Presidential Academy of National Economy and Public Administration
	Address	Moscow, Russia
	Email	malexandrov.uab@gmail.com
Author	Family Name	**Cardiff**
	Particle	
	Given Name	**John**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Institute of Technology Tallaght
	Address	Dublin, Ireland
	Email	john.cardiff@it-tallaght.ie
Author	Family Name	**Koshulko**
	Particle	

Given Name	**Oleksiy**
Prefix	
Suffix	
Role	
Division	
Organization	Glushkov Institute of Cybernetics
Address	Kyiv, Ukraine
Email	koshulko@gmail.com

Abstract

The topic of people's health has always attracted the attention of public and private structures, the patients themselves and, therefore, researchers. Social networks provide an immense amount of data for analysis of health-related issues; however it is not always the case that researchers have enough data to build sophisticated models. In the paper, we artificially create this limitation to test performance and stability of different popular algorithms on small samples of texts. There are two specificities in this research apart from the size of a sample: (a) here, instead of usual 5-star classification, we use combined classes reflecting a more practical view on medicines and treatments; (b) we consider both original and noisy data. The experiments were carried out using data extracted from the popular forum AskaPatient. For tuning parameters, GridSearchCV technique was used. The results show that in dealing with small and noisy data samples, GMDH Shell is superior to other methods. The work has a practical orientation.

Keywords (separated by '-')

Classification - Health social networks - Unbalanced data - Noise immunity - GMDH

Opinion Mining on Small and Noisy Samples of Health-Related Texts

Liliya Akhtyamova[1(✉)], Mikhail Alexandrov[2,3], John Cardiff[1], and Oleksiy Koshulko[4]

[1] Institute of Technology Tallaght, Dublin, Ireland
liliya.akhtyamova@postgrad.ittdublin.ie,
john.cardiff@it-tallaght.ie
[2] Autonomous University of Barcelona, Barcelona, Spain
malexandrov.uab@gmail.com
[3] Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia
[4] Glushkov Institute of Cybernetics, Kyiv, Ukraine
koshulko@gmail.com

Abstract. The topic of people's health has always attracted the attention of public and private structures, the patients themselves and, therefore, researchers. Social networks provide an immense amount of data for analysis of health-related issues; however it is not always the case that researchers have enough data to build sophisticated models. In the paper, we artificially create this limitation to test performance and stability of different popular algorithms on small samples of texts. There are two specificities in this research apart from the size of a sample: (a) here, instead of usual 5-star classification, we use combined classes reflecting a more practical view on medicines and treatments; (b) we consider both original and noisy data. The experiments were carried out using data extracted from the popular forum AskaPatient. For tuning parameters, GridSearchCV technique was used. The results show that in dealing with small and noisy data samples, GMDH Shell is superior to other methods. The work has a practical orientation.

AQ1

Keywords: Classification · Health social networks · Unbalanced data
Noise immunity · GMDH

1 Introduction[1]

1.1 Motivation

Social media is a modern phenomenon that has opened new possibilities for analysis of various aspects of the human society life in total or some group of peoples [1]. The medical domain is presented in various forums, where users discuss both general topics as the state of the healthcare system or the specific questions concerning medicine,

[1] Akhtyamova, L., Alexandrov, M., Cardiff, J., Koshulko, O.: Building Classifiers with GMDH for Health Social Networks (DB AskaPatient). In: Proc. of the Intern. Workshop on Inductive Modelling (IWIM-2018), IEEE, 5 pp (2018) [To be published].

N. Shakhovska and M. O. Medykovskyy (Eds.): CSIT 2018, AISC 871, pp. 1–12, 2019.
https://doi.org/10.1007/978-3-030-01069-0_27

treatment etc. Such information is of interest to various governmental and private institutions. The former has an opportunity to evaluate the reaction of community on the laws and acts concerning healthcare as well as monitor the health condition of citizens and the latter can see a market to produce medicines [2].

On the other hand, social media has provoked new developments in the Natural Language Processing (NLP) field, namely new models, methods and program systems. The medical domain presented in social media uses traditional approaches of NLP related to (1) retrieval of given cases in data, and (2) opinion mining concerning these cases. Speaking of cases, we mean specific medicines or treatments.

First, we should mention here adverse drug reactions (ADRs) which are proved to be the reason of serious injury and death of more than 700,000 people in the USA [3]. So, most of the methods developed for the analysis of health social networks are related to these ADRs. Other topics are utilizing smoking cessation patterns on Facebook [4] as well as organizing different anti-smoking and other campaigns revealing drug abuse [5] and monitoring malpractice on Twitter [6].

1.2 Problem Setting

The motivation behind this research is the consideration of limitation and noisiness of information, in this case regarding drug use. Indeed, there are often just certain users who write about problems with their health and often provide irrelevant information, pointing out their initial condition or possible the side effects of drugs, rather than their own experience. By adding noise to their reports, we reflect this issue.

In this paper we consider possibilities of GMDH-based algorithms to build useful noise-immunity classifiers for processing texts from health social networks. It is our contribution to the problem of analysis of health social media. By the term "useful classifiers", we mean classifiers which allow detecting negative or extreme cases in social media. As we mentioned above, the traditional 5-star classification includes classes = {very negative, negative, satisfactory, positive, very positive}. We denote them as {1*, 2*, 3*, 4*, 5*} respectively. Our 2-class scale includes the negative class = (1*, 2*) and the class 'others' = (3*, 4*, 5*). The 3-class scale includes the very negative class = (1*), the satisfactory class = (2*, 3*, 4*), and the very positive class = (5*). These classifications were introduced in [7]. Noise in data reduces the discriminatory between classes, therefore decreasing model accuracy. At the same time, GMDH simplifies a model to make it more stable [8]. When we speak about noise-immunity algorithms we mean here algorithms whose results are worsened less than noise grows. This worsening and growth are considered in relative units.

We intend to study various ways of text parameterization that is a transformation of the dataset to its vector form by putting attention on using not only one-word terms but also n-grams of terms and n-grams of characters.

It should be mentioned here about convolutional neural networks (CNN) having successful applications in opinion mining health social networks, see e.g. [9, 10]. However, they dealt with the traditional 5-star ratings rather than the combined classes used in this paper. They did not consider the stability of results with respect to data, and we cannot directly compare our results with those related to CNNs.

The content of this paper is as follows: Sect. 1 is the introduction, Sect. 2 describes dataset AskaPatient. In Sect. 3 we give a short description of GMDH and GMDH Shell. Section 4 presents the results of experiments with the original data. Section 5 shows the results of experiments with noisy and shortened data. We discuss our research in Sects. 6 and 7 concludes the paper.

2 Dataset

2.1 General Description of AskaPatient

The dataset AskaPatient consists of 5 fields, which are rating, the reason for taking the medication, side effects, comments, gender, age, duration and date added (AskaPatient, n.d.). As the comments usually reflect patient opinions about the drug, we left only this field for rating prediction purposes. The dataset we retrieved consists of 48,088 comments, with 32,437 left after removing duplicates. The 5-star rating distribution among comments is presented in Table 1.

Table 1. Rating distribution.

1*	2*	3*	4*	5*
12823 (27%)	5713 (12%)	8202 (17%)	9152 (19%)	12093 (25%)

With the new class distribution, the class imbalance essentially increased, as can be seen in Table 2.

Table 2. Distribution of documents on combined classes.

Contents	Class 1	Class 2	Class 3
2 classes	18536 (39%)	29449 (61%)	
3 classes	12823 (27%)	23069 (48%)	12093 (25%)

The distribution of the lengths of reviews is presented in Table 3. It could be observed that there are only 10% of reviews with a word count exceeding 200 words.

Table 3. Descriptive statistics on term count in each review.

Min number	Max number	Aver. number	90% percentile
1	823	54.4	200

For calculation simplicity, we have chosen 1,000 texts for both classification tasks, preserving the class distribution among texts. This leads to a small loss in accuracy of models, however, we were able to carry out more experiments trying different modes of the GMDH Shell platform.

2.2 Parameterization and Normalization for ML and GMDH-Based Methods

We have chosen bag-of-words (BoW) as our primary parametrization technique due to its simplicity. By choosing the best parameters for preprocessing as well as tuning models we used the GridSearchCV technique. This technique allows us to conduct an exhaustive search over specified parameter values for a classifier.

The vocabulary size varied between 100 and an unlimited number of terms. Here 'term' means word or character n-grams, where n varied between interval of [1–6].

We filtered terms which were encountered in more than 50%–75% texts. Such a limit corresponded to the first transition point with respect to the number of terms, providing discriminative power while preserving information value of obtained vectors. These vectors then were normalized to the interval [0, 1] using L2-norm. Table 4 presents all parameters that were used for tuning the BoW model.

Table 4. Word representation tuning in a grid search.

Character/word parametrization
n-gram range
Tokenization
tf-idf rate
size of dictionary

While dealing with ML methods from the scikit-learn library [11], we also tried to add a range of model-specific parameters but it did not give noticeable results. We conclude that the correct preprocessing of text data itself is more important than model specification tuning.

However, this is not the same for GMDH-based algorithms where with further model tuning in GMDH Shell platform it was possible to get the significant model improvements.

Overall, character n-grams are always superior to word ones and character n-gram in the range from 1 to 7 gives the best results. The term 'range' means here that in the process of parameterization, n-grams of different sizes are used simultaneously. Maximum dictionary size is the best option for methods in scikit-learn. Due to the computer limitations, the vocabulary size of 150 is the best option in GMDH Shell. Therefore, in all the experiments described in Sects. 4 and 5, we deal with documents presented with maximum dictionary size for ML methods in scikit-learn and in the space of dimensionality 150 for GMDH-based methods. Throughout, for ML methods character n-grams are used.

In our experiments, we used 5-fold cross-validation and weighted F-score averaged among all folds to correctly measure the model quality with unbalanced data.

2.3 Noisy Data

To form noisy data, we added an independent Gaussian noise to parameterized and normalized data with the mean = 0 and the standard deviation $s = 0.1; 0.2$.

As for the neural networks in comparison to feature-based methods they do not presume to have normalized word vectors. Indeed, as it was stated in [12] that words that are used in a similar context have longer vectors than words that are used in different contexts. Thus, the usage of raw vectors makes a model more accurate increasing its performance. Moreover, we fed to the neural networks word an embedding matrix rather than bag-of-words vectors. This is due to the fact that word2vec format embeddings give in all cases better results for neural networks than one-dimension parameterization. That is why we do not perform noise immunity analysis with neural network algorithms.

3 Methods and Tools

3.1 GMDH-Based Classifiers with Applications

Group Method of Data Handling (GMDH) is a technology of machine learning (ML) for creating noise immunity models. The ideas and applications of GMDH are presented in many publications, see for example [13–15]. Theoretical bases of GMDH are described in the well-known paper [8]. GMDH does not orient on certain class functions, but the most popular GMDH-based tools use polynomial functions of many variables [16, 17]. This fact has the simple explanation: any continuous function of many variables on hypercube can be presented in the form of uniformly-convergent polynomial series.

GMDH itself has many applications in NLP. For example, the paper [18] demonstrates the GMDH based technique for building empirical formulae to evaluate politeness, satisfaction, and competence reflecting in dialogs between passengers and Directory Enquires at a railway station in Barcelona. The formulae contain the sets of linguistic indicators preliminary assigned by experts separately for each mentioned problem (politeness, satisfaction, and competence).

The paper [19] shows the possibility of building a classifier of primary medical records using GMDH Shell. The linguistic indicators are extracted from the training dataset related to six stomach diseases. The accuracy of results on a real corpus of medical documents proved to be close to 100%. Such a result essentially exceeded the results of other methods which had been used on the same dataset.

In another paper [20], the authors present opinion classifiers for Peruvian Facebook, where users discuss the quality of various products and services. These classifiers use linguistic indicators prepared by qualified experts. The indicators form two variables reflecting the contribution of positive and negative units and then GMDH-based algorithms build polynomial models with these two variables. The total accuracy reached in the experiments significantly improved the results obtained by other researchers.

In this paper, GMDH algorithms are implemented on the platform called GMDH Shell. All algorithms related to classification realizes the One-Vs-All approach [21]

which reduces multi-class classification to the binary one. The variety of preprocessing options for this instrument could be learned from [15].

At the moment, we do not know any publications in which GMDH has been used for opinion analysis of health social networks. For this reason, it would be useful to study the possibilities of GMDH-based algorithms to classify any typical network. It would be also interesting to test the stability of results having in view the well-known property of noise immunity of GMDH-based algorithms. This paper continues our applied research presented in [22].

3.2 Standard ML Classifiers

In our experiments, we have tested several ML techniques: Random Forest, Logistic Regression, Extremely Randomized Trees, Support Vector Machine classifiers from Python scikit-learn package [11].

These tools have a long history with successful application in many research fields, e.g. Sentiment Analysis tasks. For pharmacovigilance, it was applied for example to the tasks of ADR detection [23] and monitoring prescription medical abuse on Twitter [5]. Usually, these algorithms are enriched with huge set of additional features to get better results.

3.3 Neural Networks

For comparison purposes, we included here the results of deep learning methods, however, the advantage of them is more pronounced while dealing with large data.

In this work, we construct a LSTM-CNN model for dealing with user posts. It was shown that such combined methods often achieve better results in a variety of text classification tasks [24, 25]. The intuition behind this type of networks is that output tokens from the LSTM layer store an information about not only the current token but also any previous tokens. This output of the LSTM layer is then fed to a convolutional layer which is now get enhanced information, thus making better predictions.

For preprocessing health, word embeddings were used [26]. It turned out results to be better on data with any modifications (normalization, stemming).

4 Experiments on Original Data

4.1 Experiments in GMDH Shell

Here, by original data we mean noise-free data reflected in 1,000 documents. Overall, the investigated parameters of GMDH Shell are presented in Table 5. Here *lin* denotes linear members and *sq/div.* denotes squares/divisions. The latter means the model includes linear, pairwise and square members. Complexity or rank of model means the number of features to consider which keeps some number of the most important variables according to the selected ranking algorithm. This number dramatically increases the running time of an algorithm if pairwise and square members were

included. The number of final parameters could be reduced by selecting a model complexity value.

Table 5. Options for GMDH shell tuning.

Balance	Ensemble	Form	Complexity	Rank
yes/no	yes/no	lin/sq/div.	20–200	20–300

GMDH Shell is presented in four algorithms: the combinatorial, neural network type, forward and mixed selections. The first two ones are the classical GMDH-based algorithms [14, 15]. The last two ones are the well-known algorithms of stepwise regression [27] where GMDH is used for generation of variants.

The preliminary experiments showed the following results which we considered while testing different methods of classifications:

- balancing impairs results quality;
- data transformation to different forms lead to model accuracy increase;
- ensembling, in general, leads to slightly better results;
- the model complexity, i.e. number of coefficients in a model of about size of vocabulary is always the best adjustment.
- ranking boosts model accuracy.

On the original, noise-free data mixed selection algorithm showed the best results and was chosen for the further analysis on the noisy and reduced data (Table 6). It can be observed from this table that the GMDH-based mixed selection algorithm exceeded the baseline on 32% for 2 classes and on 35% for 3 classes. Here, the baseline is denoted as the proportion of the biggest class in a classification problem, Table 2.

Table 6. F-score for different algorithms from GMDH Shell platform, original data.

Methods	2 classes	3 classes
Combi	66%	61%
Forward	82%	47%
Mixed	90%	74%
NN	61%	47%

4.2 Building Classifiers with Other Methods

The results with the best parameters are presented in Table 7.

The SVM algorithm is superior to other methods which can be explained by the fact that it is a less sophisticated algorithm, thereby less prone to overfitting. In the case of small data size that quality is essential. SVM algorithm exceeded the baseline by 15% for 2 classes and by 14% for 3 classes. Other methods gave worse results.

Table 7. F-score for different algorithms from the scikit-learn package and neural network algorithm, original data.

Methods	2 classes	3 classes
Random forest	0.63	0.41
Extra trees	0.62	0.43
SVM	0.76	0.56
Logistic regression	0.62	0.42
RCNN	0.66	0.54

For 2 classes all other ML methods gave slightly higher than baseline results. On the 3-class problem other methods did not exceed the baseline; RCNN exceeded it by 11%. However, as stated before, significant advantages of RCNN can be shown only when the sample size is tens and hundreds of thousands of documents.

5 Experiments with Noisy and Reduced Data

5.1 Building Classifiers for Noisy Data

In this sub-section, we test the noise-immunity of the best algorithm from GMDH Shell and methods from scikit-learn library. The results of the analysis in terms of rates to original means are presented in Table 8 for 2 and 3 classes accordingly.

Table 8. F-score rate for noisy data and different level of noise.

Methods	2 classes			3 classes		
	s = 0	s = 0.1	s = 0.2	s = 0	s = 0.1	s = 0.2
(GMDH) mixed	1.00	0.82	0.81	1.00	0.91	0.90
SVM	1.00	0.74	0.71	1.00	0.63	0.64
Tree-based	1.00	0.84	0.83	1.00	0.91	0.90

Tree-based and GMDH-based mixed selection methods are more stable to the noise. SVM algorithm is less prone to the noise increase stability, although outperforming tree-based methods in terms of weighted F-score.

5.2 Building Classifiers with Reduced Data

In this section, we test model performance on very small samples of data: 500 and 250 samples. This allows us to check the stability of models on the extremely small text samples. The results of the experiments for two and three classes are presented in Tables 9 and 10 accordingly.

Table 9. F-score for different ML algorithms, reduced data (500 samples).

Methods	2 classes	3 classes
(GMDH) mixed	0.92	0.79
Random forest	0.63	0.42
Extra trees	0.63	0.47
SVM	0.67	0.52
Logistic regression	0.55	0.38

Table 10. F-score for different ML algorithms, reduced data (250 samples).

Methods	2 classes	3 classes
(GMDH) mixed	0.98	0.96
Random forest	0.53	0.42
Extra trees	0.57	0.44
SVM	0.63	0.47
Logistic regression	0.46	0.34

It is noticeable that GMDH-based mixed selection algorithm is more efficient when dealing with very small data samples. Amusingly, the results for GMDH turned out to be even better with sample size reduction. The reason for this lies in the flexibility of GMDH-based algorithms which are well adjusted to the variability in the data. It is not the same for scikit-learn methods.

6 Discussion

In our paper [22], we began to study the possibilities of GMDH-based algorithms on opinion mining of typical texts related to health social networks. In the paper we conducted our experiments on the same dataset AskaPatient used in this paper. Our interest in GMDH as a technology of text mining was provoked by the following circumstances: GMDH can successfully deal with small amount of experimental data; moreover, it works well even when the dataset size is less than the number of parameters used; GMDH builds models of optimal complexity that provide their high noise immunity. With these circumstances, our study of GMDH-based algorithms was quite limited: we did not consider the sensibility of models to size of experimental data and we did not consider the noise-immunity of models built.

GMDH-based classifiers are not the only ones that can be used for opinion mining. Last year the great popularity came to program language Python and tools based on it. This fact provokes comparison of classifiers built on GMDH technology [13–17] and classifiers included in the well-known Python library scikit-learn [11].

In the current research, we tried to explore all mentioned problems by putting special attention to parameter tuning, in particular, experiments with different type of parametrization: character or word n-grams. In the paper [22] we used only one-word

terms. To select these terms, we used the criterion of term specificity which considers term frequency in a given document corpus and any basic corpus [28, 29].

In that research, we used word frequency list related to British National Corpus as this basic corpus. In the current research, we carefully studied different combinations of n-grams of terms and n-grams of characters to select the best parameters. Such a process is described in the Sect. 2.2. Table 11 shows results of classification for n-grams of terms and n-grams of characters. We studied also results of classification related to different number of posts where a given term occurs. The results are presented in Table 12.

Table 11. Study of different sizes of vocabularies.

Options	Sizes	Results
n-grams of characters	50, 150, 250, 400	150–250 give the best and close results
n-grams of terms	150, 250, 400, 800	150–250 give the best and close results

Table 12. Study of different number of posts.

Option	Number of posts	Results
Posts with a given term	>25%, >50%, >75%	75% gives the best results

The results presented above defined options which we used in this paper.

7 Conclusions

In the paper, we investigated the noise-immunity and data size sensitivity of different algorithms on health-related texts. It was stated that user reports on drugs are good examples of very noisy data where it is often that the information is quite limited on some drugs and especially their side effects. Thus, while dealing with imbalance it is needed to deal with small samples of text and noise in data. For these purposes, we built different machine learning classifiers including standard machine learning classifiers as well as GMDH-based algorithms and neural networks.

We tested different preprocessing options and found out that character n-grams with absent lemmatization and stemming work the best in all cases. Overall, GMDH-based mixed selection algorithm performs better on small and extremely small text samples. Moreover, it is more stable to adding a noise in comparison to the standard ML methods. This might be explained by the fact of more simplicity and flexibility of the GMDH-based algorithms in comparison to tree-based and SVM algorithms. The results have clear practical implications and can be used in further research.

References

1. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of Social Media (2007). https://doi.org/10.1016/j.bushor.2009.09.003
2. Ventola, C.L.: Social media and health care professionals: benefits, risks, and best practices. P T **39**, 491–520 (2014)
3. Lehne, R.A., Rosenthal, L.D.: Pharmacology for Nursing Care. Elsevier Health Sciences (2013)
4. Struik, L.L., Baskerville, N.B.: The role of Facebook in crush the crave, a mobile- and social media-based smoking cessation intervention: qualitative framework analysis of posts. J. Med Int. Res. **16**(7), e170 (2014). https://doi.org/10.2196/jmir.3189
5. Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., Gonzalez, G.: Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. Drug Saf. **39**, 231–240 (2016)
6. Nakhasi, A., Passarella, R.J., Bell, S.J., Paul, M.J., Dredze, M., Pronovost P.J.: Malpractice and Malcontent: analyzing medical complaints in Twitter. In: AAAI Technical Report FS-12-05, Information Retrieval and Knowledge Discovery in Biomedical Text, pp. 84–85 (2012)
7. Alexandrov, M., Skitalinskaya, G., Cardiff, J., Koshulko, O., Shushkevich, E.: Classifiers for Yelp-reviews based on GMDH-algorithms. In: Proceedings of the Conference in Intelligent Text Processing and Comput. Linguistics (CICLing-2018). LNCS, pp. 1–18. Springer (2018)
8. Stepashko, V.S.: Method of critical variances as analytical tool of theory of inductive modeling. J. Autom. Inf. Sci. **40**, 4–22 (2008). https://doi.org/10.1615/J.AutomatInfScien.v40.i3.20
9. Huynh, T., He, Y., Willis, A., Uger, S.: Adverse drug reaction classification with deep neural networks. In: Proceedings of 26-th International Conference on Computational Linguistics (COLING-2016), pp. 877–887 (2016)
10. Akhtyamova, L., Ignatov, A., Cardiff, J.: A Large-scale CNN ensemble for medication safety analysis. In: Proceedings of 22th International Conference on Applications of Natural Language to Information Systems (NLDB 2017). LNCS, pp. 1–6. Springer (2017)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
12. Schakel, A.M.J., Wilson, B.J.: Measuring word significance using distributed representations of words, CoRR, abs/1508.02297 (2015)
13. Madala, H.R., Ivakhnenko, A.G.: Inductive Learning Algorithms for Complex Systems Modelling. CRC Press, New York (1994)
14. Farlow, S.J.: Self-Organizing methods in modeling: GMDH type algorithms. In: Statistics: A Series of Textbooks and Monographs, Book 54, 1-st edn. Marcel Decker Inc., New York, Basel (1984)
15. Stepashko, V.: Developments and prospects of GMDH-based inductive modeling. In: Shakhovska, N., Stepashko, V. (eds.) Advances in Intelligent Systems and Computing II / AISC book series, vol. 689, pp. 346–360. Springer, Cham (2017)
16. Platform GMDH Shell. www.gmdhshell.com
17. Resource GMDH in IRTC ITS NAS of Ukraine. mgua.irtc.org.ua/
18. Alexandrov, M., Blanco, X., Catena, A., Ponomareva, N.: Inductive modeling in subjectivity/sentiment analysis (case study: dialog processing). In: Proceedings of 3-rd International Workshop on Inductive Modeling (IWIM-2009), pp. 40–43 (2009)

19. Kaurova, O., Alexandrov, M., Koshulko, O.: Classifiers of medical records presented in free text form (GMDH shell application). In: Proceedings of 4-th International Conference on Inductive Modeling (ICIM-2013), pp. 273–278 (2013)
20. Alexandrov, M., Danilova, V., Koshulko, A., Tejada, J.: Models for opinion classification of blogs taken from Peruvian Facebook. In: Proceedings of 4-th International Conference on Inductive Modeling, pp. 241–246 (2013)
21. Tax, D.M.J., Duin, R.P.W.: Using two-class classifiers for multiclass classification. In: Proceedings of 16-th International Conference on Pattern Recognition, pp. 1051–1054. IEEE (2002)
22. Akhtyamova, L., Alexandrov, M., Cardiff, J., Koshulko, O.: Building classifiers with GMDH for health social networks (DB AskaPatient). In: Proceedings of the International Workshop on Inductive Modelling (IWIM-2018). IEEE (2018). [to be published]
23. Sarker, A., Gonzalez, G.: Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J. Biomed. Inform. **53**, 196–207 (2015). https://doi.org/10.1016/j.jbi.2014.11.002
24. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Proceedings of 16th International Conference on Artificial Intelligence, pp. 2266–2273 (2015)
25. Stojanovski, D., Strezoski, G., Madjarov, G., Dimitrovski, I.: Finki at SemEval-2016 Task 4: deep learning architecture for Twitter sentiment analysis. In: Proceedings of SemEval-2016, pp. 149–154 (2016)
26. Miftahutdinov, Z., Tutubalina, E., Tropsha, A.: Identifying disease-related expressions in reviews using conditional random fields. In: Proceedings of International Conference on Computational Linguistics and Intellectual Technologies (Dialog-2017), pp. 155–166 (2017)
27. Draper, N., Smith, H.: Applied Regression Analysis. Wiley, New York (1981)
28. Gelbukh, A., Sidorov, G., Lavin-Villa E., Chanova-Hernandez, L.: Automatic term extraction using Log-likelihood based comparison with General Reference Corpus. In: Proceedings of 15-th International Conference on Applications of Natural Language to Information Systems (NLDB-2010). LNCS, vol. 6177, pp. 248–255. Springer (2010)
29. Lopez, R., Alexandrov, M., Barreda, D., Tejada, J.: LexisTerm – the program for term selection by the criterion of specificity. In: Artificial Intelligence Application to Business and Engineering Domain, vol. 24, pp. 8–15. ITHEA Publ., Rzeszov-Sofia (2011)

Author Query Form

Book ID : **471168_1_En**

Chapter No : **27**

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

Dear Author,
During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	Please confirm if the corresponding author is correctly identified. Amend if necessary.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	New matter followed by
Insert in text the matter indicated in the margin		or (x2)
Delete	through single character, rule or underline or through all characters to be deleted	or (x2)
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	