

Classifiers for Yelp-reviews based on GMDH-algorithms

Mikhail Alexandrov^{1,2}, Gabriella Skitalinskaya³, John Cardiff³,
Olexiy Koshulko⁴, Elena Shushkevich³

¹Russian Presidential Academy of National Economy and Public Administration,
Moscow, Russia

²Autonomous University of Barcelona, Barcelona, Spain

³Institute of Technology Tallaght, Dublin, Ireland

⁴Glushkov Institute of Cybernetics, Kyev, Ukraine

MAlexandrov@mail.ru, gabriellasky@icloud.com,
john.cardiff@it-tallaght.ie, koshulko@gmail.com,
e.shushkevich@yandex.ru

Abstract. Yelp is one of the most popular international web resources about products and services that provide users with useful information on local businesses and helps the business owners to make their business more attractive for the users. The Yelp dataset consists of attributes for describing the business, reviews in free text form and numeric star ratings out of 5. The utility of such a dataset has provoked dozens of publications related to classifiers of ratings, which used various smart tools of opinion mining. Unlike them, in this paper we propose to use simpler approaches, namely: (a) selection of descriptors based on term specificity, and (b) formation of classifiers with these descriptors based on inductive modeling. The latter is implemented by the well-known tool GMDH Shell, where GMDH stands for Group Method of Data Handling. This method allows us to build models with high noise immunity. We compare 96 prediction models with identified descriptors by combining various variants: (i) preprocessing with data transformation and balancing classes, (ii) algorithms of classification; and (iii) post processing with ensembling. Instead of the typical 5- star classification we consider combined classes reflecting more practical view on purchase of goods or development of business. The experiments refer to the most popular categories of business: restaurants and shopping. To evaluate the quality of classifiers we consider the results of predecessors, and we also introduce the so-called defensible accuracy. With this comparison the results presented in the paper prove to be promising.

Keywords: Yelp, GMDH, GMDH Shell, text mining, opinion mining

1 Introduction

1.1 Motivation

adfa, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011

Our motivation for undertaking this research is to answer several questions, namely: *why does the study of forums like Amazon, Epinions and Yelp prove to be important for business development? why do owners of business take into account these forums? why do users prefer post their notes about products and services?*

Numerous studies (see, e.g. [5]) clearly show that forums have a significant impact on consumer purchase decisions as well as on business revenues. The same studies also show that there are no direct relations between a given review and a given star rating. So, one needs to generalize the existing information using dozens or even hundreds reviews concerning a given business or similar businesses. Obviously, this procedure is time consuming and it needs application of effective computer tools.

Many small business owners are motivated to start their own business not only because they want to have any material benefits, but also because they want to change their way of life, as well as other personal factors. The study [39] shows that non-financial objectives can lead to alternative measures of success, especially in small business sector. This trend is based on the fact that all financial characteristics indirectly imply that company would like to grow and to increase their capital. However, some companies are not interested in growth and it means that financial indicators such as profit are not their main and the only motivation. Therefore here business owners have other non-financial criteria to measure their success. One of such factors is the personal satisfaction of these business owners. The personal satisfaction and achievements along with the pride for work and a free way of life are often valued higher than material goods.

Here we can ask a question about the other opportunities to know client opinions. For example, these companies could use their own sites. This situation was studied in [13] with the following conclusions: socially-oriented sites are considered trustworthy, and the opinions presented in these sites are considered impartial. The effect of trust in socially-oriented sites has a strong influence on the involvement of users to on-line activities, while non-social sites do not have a significant impact.

Users also have strong motivation to communicate their opinions on the sites related to products and services. We could mention here the following motives [9]: a) a consumer wants to contribute to a community by posting his/her own review and comments on products and services being interesting to other members of the community; b) a consumer feels satisfaction when other participants of a given Internet platform approve his/her contribution (such a feedback can be formal from platform operators or informal from other users); c) a consumer has a complaint against a certain company, and the Internet platform facilitates the presentation of complaints (it is possible if there is a third party - the moderator of the system, which communicates with the company on behalf of the client).

1.2 Related work

1. Opinion mining Yelp dataset

The Yelp dataset [40] contains information about approximately 1 million businesses. Each business is described: a) formally by means of its attributes reflecting its location and functionality; and b) informally on the basis of reviews and

star assessments. This dataset is used in numerous applications concerning business of products and services. We consider only those related to text processing.

The Yelp dataset contains many omitted and noisy data. For this cases, the authors of [8] propose a generic approach to factorization of data that jointly models relations in the Yelp database. Here ‘data’ and ‘relations’ mean reviews, attributes, and categories of business. Having revealed a set of factors that are shared across existing data the model is able to reflect the information about other relations. The paper also presents joint visualizations of factors being built on terms, attributes and categories and shows some dependencies between them that are not directly observed in the data.

The authors of [25] also deal with hidden factors but in this case, they reveal them in the framework of the problem of personalization. The authors build hidden factors and join them to hidden topics, which are revealed using LDA (Latent Dirichlet Analysis). This approach allows yields: a) easy interpretable textual labels for latent rating dimensions, which helps to ‘justify’ ratings with text; and b) discovered topics that can be used to facilitate solutions to other problems related to automated genre determination, and to identify useful and representative reviews. The examples refer to Amazon and Yelp collections.

Traditional topic modeling lacks methods of incorporating star ratings or semantic analysis in the generative process. The author of [20] proposes a modified LDA, in which term distributions of topics are conditional on star ratings. It is easy to see that such an approach reflects personalization of solutions. The author shows that where one examines topic mixture in documents then this approach produces clearer and more semantically-oriented topics than those of traditional LDA. The experiments use the Yelp dataset [40].

The paper [21] refers to the problem of text annotation. The authors propose a new method for extracting quality phrases from text corpora integrated with phrasal segmentation. This method requires only limited training but the quality of generated phrases proves to be close to human judgment. The method is scalable: both computation time and required space grow linearly as the corpus size increases. The experiments are performed on Academia and Yelp collections.

Yelp restaurant reviews is the subject of consideration in [10]. The authors use on-line LDA to discover latent subtopics on the basis of a large amount of reviews with high dimensionality. These subtopics can provide meaningful insights to restaurants about needs of customers in order to increase their Yelp ratings, which directly affects the revenue of restaurant owners. The paper shows the breakdown of hidden topics over all reviews, predicts stars per hidden topics discovered, and extends our findings to that of temporal information regarding restaurants peak hours.

The authors of the recently published paper [3] attack the problem of predicting a user’s star rating from two parts: a) extraction of the best set of features from given texts, and b) choice of the best machine learning algorithm. The former uses 4 feature extraction methods: (i) unigrams, (ii) bigrams, (iii) trigrams, and (iv) latent semantic indexing. The latter uses 4 machine learning algorithms: (i) logistic regression, (ii) Naive Bayes classification, (iii) perceptrons, and (iv) linear Support Vector

Classification. Taken together these variants form 16 models for predicting the ratings from reviews. The authors analyze the performance of each of these models on Yelp dataset and propose the best one. This publication gives a good baseline for those who deal with various problems of classifications on Yelp dataset.

2. Classification of texts with GMDH

GMDH (Group Method of Data Handling) is a method of machine learning, which allows us to build models of optimal complexity from a given class of models to describe experimental data [7, 26, 33]. Due to its universality the polynomials of many variables prove to be the most popular class in many applications. In particular, this class is used in the well-known tool GMDH Shell [30] and also in the majority of tools presented in [31]. Hereinafter we consider only GMDH applications related to classification of textual data.

The paper [1] demonstrates the application of the GMDH based technique for building empirical formulae to evaluate politeness, satisfaction and competence reflecting in dialogs between passengers and Directory Inquires of a railway station in Barcelona. The formulae contain sets of linguistic indicators preliminary assigned by experts separately for each mentioned problems (politeness, satisfaction and competence).

In [2], the authors present the results of building opinion classifiers for Peruvian Facebook, where users discuss the quality of various products and services. The authors use a) linguistic indicators prepared by experts, which automatically form two variables that determine the contribution of positive and negative units, and b) GMDH Shell tool [28] for the selection of optimal model in the class of polynomial models including the mentioned variables. Each formula reflects the contribution of positive/negative units in a text. The total accuracy reached in the experiments significantly improved the results obtained by other researchers.

The paper [15] demonstrates possibility of building a classifier of primary medical records using GMDH Shell. The linguistic indicators are extracted from training data sets 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 data set. In this paper, one builds classifiers of texts reflecting opinions of currency market analysts about euro/dollar rate. The classifiers use various combinations of classes: growth, fall, constancy, not-growth, not-fall. The process includes term selection based on criterion of term specificity and model selection using technique of inductive modeling. The latter is implemented with GMDH Shell tool mentioned above. The experiments evaluate quality of classifiers and their sensibility to term list. This work has an essential practical orientation.

1.3 Problem setting

The review presented above defines the contents of the paper: we study possibilities to predict star rating using less smart and more interpretable methods. To select terms we use the procedure based on criterion of term specificity. Such a criterion compares the relative frequency of term occurrence in a given corpus and in

some standard corpus. To build the model itself we use the Group Method of Data Handling (GMDH). This method (it is better to say ‘technology’) evaluates step by step models of a given class from the simplest ones to the more complex ones to provide the best noise immunity. Neither criterion of term specificity nor GMDH have been used before in the problem under consideration.

Unlike the typical 5 star rating for 5 categories of success we consider classification on 2 classes and 3 classes; each of them is the certain combination of several star categories. These classifications are easy interpretable and allow to take into account extreme classes and relatively successful classes.

We test the values of the mentioned criterion of specificity and select a compromise between the limited and redundant term lists. Here we take into account that the less informative terms will be then automatically filtered by GMDH. To build the best model we combine different options of preprocessing and post processing, and also test several methods of classification. The best model is selected separately for grouping on 2 classes and 3 classes. Totally we deal with 120 variants. Term selection and all procedures of classification are completed by the program LexisTerm [22] and the package GMDH Shell [30].

To evaluate the results of experiments we take into account not only the results of previous research, but also the so-called defensible accuracy. We introduce this notion to take into account the coincidence of expert opinions concerning star rating. Naturally, if experts have different opinions then we should not try to reach the accuracy of 100%. Unfortunately, this circumstance is not taken into account in applied research with subjective human opinions.

In the experiments we use two datasets of 1000 objects each taken from the Yelp resource [40]. The first one is related to restaurants and the second one is related to shopping. Restaurants and shopping are the most popular topics in Yelp. These two samples are the representative ones from the point of view of object distribution between different star categories.

The paper is organized as follows. Section 2 describes linguistic resources. Section 3 introduces classifications used in the research. Section 4 presents the technique of modeling. Section 5 contains the results of experiments. Section 6 concludes the paper.

2 Data description

2.1 Representativity of samples

The Yelp dataset contains approximately 4 million reviews about activity of 0.1 million companies. These companies reflect about 1,000 types of business. The general characteristics of Yelp dataset presented in Table 1 relate to the 5 most numerically popular business categories [3]. It should be noted that some reviews reflect opinion about several aspects of the same business simultaneously (e.g. shopping and food). For this reason the total % of reviews exceeds 100% of these 5 business categories.

Table 1. Distributions of companies and reviews [%]

<i>No.</i>	<i>Categories</i>	<i>Companies</i>	<i>Reviews</i>
1	Restaurants	34	68
2	Shopping	15	6
3	Food	12	13
4	Home services	12	10
5	Beauty	8	4

In the experiments we consider reviews related to restaurants and shopping as they are the most popular. The selection of 1,000 documents was completed by a random way. To test the representativity of the selected data we compared the star distributions for the category Restaurants based on Yelp [3] with our sample. The results are presented in Table 2. One can see the closeness of both distributions, so our sample can be assumed to be representative.

Table 2. Star distributions [%]

<i>Dataset</i>	<i>5*</i>	<i>4*</i>	<i>3*</i>	<i>2*</i>	<i>1*</i>
Yelp (restaurants)	33	33	16	10	8
Sample (restaurants)	27	34	19	10	10
Sample (shopping)	34	28	13	9	16

We have no data concerning the star distribution for the category Shopping based on Yelp. To avoid this difficulty we compare the star distributions for restaurants and shopping on our samples. These distributions are expected to be similar due to the similar needs of users with respect to these services. Table 2 shows the closeness of these distributions and this circumstance can be an indirect proof that the sample for the category Shopping can be also assumed to be a representative sample.

2.2 Term selection

1. Two approaches to term selection

The review-based rating prediction is a problem of sentiment analysis, which relies on different descriptors extracted from a given text/corpus. By ‘descriptor’, we mean a term or term combination, whose frequencies are used in predicting model. There are two different approaches used for descriptor selection: lexical (lexicon-based) approach [38] and machine learning approach [28].

The lexical approach is based on semantic orientation (SO) lexicons (words with their semantic orientation) and calculates an overall sentiment by aggregating the values of those words presented in a text or a sentence. The classical example of lexical approach is the tool SO-Calc widely used in many applications. Its vocabularies contain approximately 5,500 terms distributed between 4 vocabularies (nouns, verbs, adverbs, and adjectives together with intensifiers) [37]. The integer SO value assigned to each term varies between -5 and 5 and the sum of these values determines the polarity of opinion of a document under consideration. The SO-Calc vocabulary was tested in [14] using various scales of SO: between -3 and 3, between -2 and 2, and the binary one {-1,1}. With the binary scale the authors reached accuracy

0.78 vs 0.83 with the SO-Calc vocabulary. It means that it is not necessary to use too fine-grained scales.

The machine learning approach uses collections of labeled texts as a training data in order to build automated classifiers. The machine learning approach is presented in the well-known comprehensive survey [29]. The majority of cases presented in this survey concern binary classification. The authors working with rating prediction use more smart ways. For example, in [19] one builds a vocabulary of sentiments using term strength with respect to each of 5 classes. The term strength is determined by the relative frequencies of term occurrence in these classes. Then this vocabulary is used in collaborative filtering algorithms.

2. Criterion of term specificity

In our work we use combined approach consisting of two phases. Initially we build a domain-oriented lexicon using criterion of term specificity. Obviously this phase refers to a lexical approach. Then we select the most informative terms in the process of inductive modeling. The latter is one of the technologies of machine learning.

By ‘Term Specificity’ with respect to a given corpus we mean a factor $K \geq 1$, which shows how much term frequency in the corpus $f_C(w)$ exceeds its frequency in any standard corpus $f_L(w)$: $K = f_C(w) / f_L(w)$. In our work we use the General Lexis of English that reflects term frequencies in the British National Corpus. This lexis is available online [16]. It is easy to see that the higher K is, the less number of terms is selected. The experience shows that in all cases when $K \geq 3$ then both stop words and almost all general lexis are eliminated. So, users may not to think about this kind of words.

We built lists of terms for $K=2,5,10,20,50$ separately for the categories Restaurants and Shopping, in total 8 lists. Then the clearly unuseful words were removed from each list. We used here the logarithmic step for variation of K to obtain essential changes in these lists. Table 3 shows the sizes of some lists after and before correction (in parenthesis). All examples are presented as stems. One can see that when the factor K changes on logarithmic scale the size of lists changes in lineal (almost lineal) scale. It is typical for different domain-oriented corpus of documents.

It was found that: when $K=2$ the lists included many insignificant terms, when $K>5$ we lost many useful terms. So, we took the threshold $K=5$ as the compromise. Table 3 shows the sizes of some lists after and before correction (in parenthesis). All examples are presented as stems

Table 3. Sizes of lexicons

<i>Dataset</i>	<i>K=5</i>	<i>K=10</i>	<i>K=20</i>	<i>Examples (stems)</i>
Restaurants	120 (142)	68 (70)	33 (33)	amaz beef bowl cool ...
Shopping	127 (160)	41 (46)	17 (18)	amaz bike brand dress ...

The criterion of term specificity is realized in the program *LexisTerm*, which is described in details in [20]. This program was used in our projects in Spain, Peru and Russia [2,15,17]. We also used it in this research.

It should be noted that the criterion of term specificity proves to be very useful not only for analysis of Yelp reviews. It can be also useful for many other applications, where topic-oriented vocabularies are necessary for numerical presentation of documents. The typical way for building such vocabularies with the mentioned criterion consists of standard steps:

- 1) Experts in a given area present a representative set of documents related to this area. Speaking of the total volume of documents (number of their words) one should take into account one remarkable regularity, namely: the number of new terms increases in arithmetic progression, when the total number of terms increases in geometric progression. This lexical regularity is firstly described in [36].
- 2) Computer linguist builds a topic oriented vocabulary using the criterion of term specificity. He/she has possibility to manage the deepness of topic presentation by changing the factor K . Therefore we have here the scalable method of topic presentation.

By the way if one knows in advance the approximate number of subtopics in the framework of a given topic (from... to ...) then it is possible to build vocabularies for these subtopics simultaneously with the vocabulary for the whole topic. For this it is necessary to use the criterion of term specificity and technique of clustering. This method is described in [24].

2.3 Parameterized documents

1,000 documents of the category Restaurants and 1,000 documents of the category Shopping mentioned above are presented in vectorial form in the space of selected terms. Therefore we have two matrices document-term containing term frequencies. The characteristics of matrices are shown in Table 4. Here: 'Dimension' reflects the number of documents and terms, 'Max freq.' is equal to the maximum number of term occurrences, 'Zero %%' shows the percent of zero values in matrices. The other data are the number of documents having a given rank. Obviously, the values 1* and 5* mean the worst and the best points.

Table 4. Characteristics of matrices

<i>Dataset</i>	<i>Dimension</i>	<i>Max freq.</i>	<i>Zero %%</i>	<i>No. 5*</i>	<i>No. 4*</i>	<i>No. 3*</i>	<i>No. 2*</i>	<i>No. 1*</i>
Restaurants	1000 x 120	18	90	267	344	193	97	99
Shopping	1000 x 127	14	97	346	276	133	88	157

We also measured the completeness of document images, that is the number of key-terms the images include. The results are presented in Table 5. Here: 'Aver. number' is average number of terms in documents of a corpus under consideration.

Table 5. Number of key-terms in documents

<i>Dataset</i>	<i>Min number</i>	<i>Max number</i>	<i>Aver. Number</i>
Restaurants	1	91	15.6
Shopping	1	69	9.9

These tables show that the matrices are very sparse that worsens the results of classification. For this reason in our next research we plan to enrich the lists of terms with the most significant 2-grams and 3-grams of letters. For selection of these k -grams the same criterion of term specificity is supposed to be used.

3. Classifications and its quality

3.1 Combined classes

The assessment of a business in Yelp-reviews is evaluated by means of ‘stars’. They reflect rank classification from 1* to 5*. For this reason the quality of automatic classifiers in publications is also evaluated on this scale (see, for example [3,8]). From the other hand, such a 5-class scale is seemed to be too detailed for real applications, when it is necessary only to say whether a given business is successful/unsuccessful or whether this business is extreme/satisfactory. For this reason, in this paper we use combined classes that are better oriented on practical situations. The corresponding classifications are presented in the Table 6 using the 5-class scale.

Table 6. Combined classes

<i>Contents</i>	<i>Success of companies</i>	<i>Class 1</i>	<i>Class 2</i>	<i>Class 3</i>
2 classes	unsuccessful, others	1*, 2*	3*, 4*, 5*	
3 classes	failed, satisfactory, excellent	1*	2*, 3*, 4*	5*

With the data from Tables 4 and 6, we can calculate new distributions of documents between combined classes. They are presented in Tables 7 and 8. It is easy to see that the distribution between combined classes proves to be essentially more unbalanced in comparison with the distribution between initial star-classes. So, we carefully tested the option of balancing classes in the process of modeling (see section 5)

Table 7. Distribution of documents on 2 combined classes (unsuccessful-other)

<i>Dataset</i>	<i>Class 1 (1*,2*)</i>	<i>Class 2 (3*,4*,5*)</i>
Restaurants	196 (20%)	804 (80%)
Shopping	245 (25%)	755 (75%)

Table 8. Distribution of documents on 3 combined classes (failed-satisfactory-excellent)

<i>Dataset</i>	<i>Class 1 (1*)</i>	<i>Class 2 (2*,3*,4*)</i>	<i>Class 3 (5*)</i>
Restaurants	99 (10%)	634 (63%)	267 (27%)
Shopping	157 (16%)	497 (50%)	346 (34%)

The proposed classification on 2 combined classes may be useful when one needs to avoid unsuccessful purchase or unsuccessful development of his/her business. The proposed classification on 3 combined classes may be useful when we are ready to

purchase something with satisfactory quality or to remain with satisfactory level of business having avoided the extreme situations.

3.2 Defensible accuracy

In the paper we propose an approach for building classifiers related to Yelp-reviews. These classifiers are tuned to the new classes more oriented on practical needs of customers and businessmen. The question is how to evaluate the quality of classifiers and how to evaluate their advantage?

The quality of classifiers can be assessed with the different measures described in the well-known books on Information Retrieval [4,23]. The typical approach is a comparison of results of classification with a given Gold Standard (GS). The most popular measure here is so-called group F -measure introduced in the paper [32]. This measure takes into account the values of recall and precision for each class and the sizes of these classes. In case of classification (unlike clustering) it can be written as follows:

$$F = \sum_i (n_i/N) \max_j (F_{ij})$$

Here: $i = \{1, 2, \dots, m\}$ is the counter for classes in GS; $j = \{1, 2, \dots, k\}$ is the counter for classified groups (clusters); m is the number of classes; k is the number of groups (clusters); n_i is the number of documents in class i ; N is the total number of documents; F_{ij} is a partial F -measure for group j with respect to class i .

The other well-known measure is so-called A -measure, that is accuracy. It is very simple and essentially less smart measure than F -measure. Usually accuracy takes into account the relative number of successful cases of classification independently of classes. But we use weighted accuracy as more adequate measure taking into account the successful cases in each class:

$$A = \sum_i (n_i/N) A_i$$

Here: $i = \{1, 2, \dots, m\}$ is the counter of classes in GS; m is the number of classes; n_i is the number of successfully classified objects from the class i ; N is the total number of objects in the dataset; A_i is a partial A -measure for the class i .

Our past experience shows that in case of hundreds or even dozens of objects and a small number of classes the measures F and A prove to be approximately equal. So, we will use A -measure having in view this circumstance.

To evaluate the advantage of the proposed classifiers we should compare our results with the others obtained before us. But here we meet the principal difficulty: our classifiers use combined classes described above but all our predecessors dealt with the 5-star classification. In order to cover such a problem we introduce so-called *defensible accuracy*, which shows the upper level of accuracy to be reached by the method under consideration. The fact is we consider opinions taken from social networks. The authors of blogs and posts from social networks are only users but not experts in the area. So, their opinions can't be accepted as the final truth. Moreover even experts in the area often have different opinions concerning the same subject or event although the difference between their opinions has less variation. Therefore,

when we want to assess any classifiers for social networking then we need to take into account this uncertainty.

To evaluate the defensible accuracy we need:

1. To assess the same set of reviews by several experts
2. To assess the concordance between experts

The relative number of fully coincident assessments just defines the defensible accuracy. Such a term means that any accuracy higher than the defensible accuracy has no sense because experts have different opinions with respect to given object or event. In our case these objects or events are given reviews.

The concordance of expert's opinions means the consistency of a given group of experts. If this group is not concordant then the obtained defensible accuracy is unreliable.

For the experiments we selected 50 reviews of the category Restaurants and 50 reviews of the category Shopping. The star distribution in each mini datasets corresponded to star distribution in the entire dataset of 1,000 reviews reflected in Table 4. These reviews were additionally evaluated by two experts with native English (co-authors of the paper). As an example of data processing we show assessments concerning category Restaurants in Table 9. Here: 'Yelp' is rating according Yelp dataset, 'Exp1' and 'Exp2' are ratings of two experts mentioned above, 'Class' is combined class according 2-class grouping, 'Equality: Rating / Class' reflects the coincidence of opinions about rating and class among the experts. Class A is equal {3*,4*,5*}, class B is equal {1*,2*}. Coincidence is marked by '1' in case of full coincidence and '0' in other cases.

Table 9. Assessments for documents concerning restaurants (examples)

<i>Examples of documents</i>	<i>Yelp / Class</i>	<i>Exp1 / Class</i>	<i>Exp2 / Class</i>	<i>Equality: Rating / Class</i>
Good selection of different poutines but otherwise nothing special. Fries weren't the greatest and couldn't taste the pepper sauce. Plus is that it's open 24h for those times when you got the munchies post bar hopping.	2 / B	2 / B	2 / B	1 / 1
It's an always open creative greasy spoon, popular with after hours crowds who are jonesing for late night eats. There aren't many 24/7 places, and this one is a Montreal staple. Personally, I'd rather eat a bag of Doritos.	2 / B	1 / B	3 / A	0 / 0
Food was by far the worst tasting Mexican 'food' I've ever had. The chips and salsa were flavorless and my burrito had noodles in it! Topped off by the fact that I immediately got sick after eating the vegetarian burrito. I would not recommend this establishment. I can't believe they're still open for business.	1 / B	1 / B	1 / B	1 / 1

To calculate the coincidence of expert opinions we used 20 reviews from the mentioned 50 reviews for the work with 2 classes, and the other 30 reviews for the work with 3 classes. The results are presented in Table 10. The data from this Table can be considered as the upper level of the defensible accuracy.

Table 10. Coincidence of expert opinions (3 experts)

<i>Category</i>	<i>2 classes</i>	<i>3 classes</i>	<i>5 classes</i>
Restaurants	95%	83%	68%
Shopping	97%	85%	64%

To check the concordance of experts we applied the criterion of Kendall that is the criterion of rank correlation. The Table 11 contains the values of this criterion. The 5%-threshold related to Kendall criterion is equal $T=0.32$ for 20 reviews and $T=0.25$ for 30 reviews. Because all values exceeds these thresholds then Yelp-expert, Expert-1 and Expert-2 belong to the one coordinated group of experts, which we can trust to.

Table 11. Values of Kendall coefficient of concordance

<i>Category</i>	<i>Experts 1-2 (2 classes)</i>	<i>Experts 1-3 (2 classes)</i>	<i>Experts 2-3 (2 classes)</i>	<i>Experts 1-2 (3 classes)</i>	<i>Experts 1-3 (3 classes)</i>	<i>Experts 2-3 (3 classes)</i>
Restaurants	1.00	0.99	0.99	0.92	0.88	0.96
Shopping	1.00	1.00	1.00	0.95	0.93	0.93

4. Method and tools

4.1 Group Method of Data Handling (GMDH)

To build classifiers we use technique of inductive modeling presented by the Group Method of Data Handling (GMDH), developed by the remarkable Ukrainian scientist O. Ivakhnenko. His first International publications related to GMDH appeared in the 1970s [11,12], but in spite of the long history of GMDH it is still not well-known to researchers. So, we give here a brief description of GMDH:

1. An expert defines a sequence of models, from the simplest to more complex ones.
2. Experimental data are divided into two datasets: training data and control data
3. For a given kind of model, the best parameters are determined with training data using any internal criterion
4. This model is tested on control data using external criteria
5. The external criteria (or the most important one) are checked on having reached a stable optimum. In this case the search is finished. Otherwise, a more complex model is considered and the process is repeated from step 3.

Naturally, this description is a basic one without details about a partition of dataset on subsets, or a process of search, or criteria. For example, often the dataset is divided on three datasets: training data, control data, and exam data. The latter is used for testing quality of selected model having in view that control data is only the filter

for model selection. Besides, sometimes the search is organized in two directions simultaneously: from the simplest model to more complex ones and from the most complex model to the simpler ones, and so on. These details are reflected in recently published work [6]. The full survey concerning the history and perspectives of GMDH is published in [33]. The theoretical basis for GMDH approach is presented in [34].

The typical form of model presentation is the polynomial one:

$$y = a_0 + \sum a_i x_i + \sum b_{ij} x_i x_j + \sum c_{ijk} x_i x_j x_k + \dots$$

Here: y is a dependent variable, x_i are independent variables, a, b, c are coefficients to be determined. We can use both positive and negative power functions x^m , where $m < 0$, or $m > 0$. If we consider the process on half-infinite or infinite intervals then instead of presented polynomial we may use the systems of classical orthogonal polynomials of Laguerre and Hermite respectively.

GMDH selects models of optimal complexity in the framework of a given class of models and it is the principal property of GMDH technique. By ‘complexity’ we mean the number of parameters for model description. Optimal complexity provides the best noise immunity of models: when noise increases the model automatically becomes simpler and vice versa. Such an effect is considered in [34].

One can find the full information about GMDH development and applications in [31].

4.2 GMDH Shell

GMDH Shell (GMDH-S) is a well-known tool for the following applications:

- time series prognosis (extrapolation),
- function presentation (approximation),
- object classification

including extended possibilities for visualization of results [30]. GMDH-S employs the technique of GMDH. At present GMDH-S includes 2 classical algorithms with their modifications:

- Combinatorial GMDH,
- GMDH-type neural networks.

In our research we use the classification option. For this GMDH-S uses One-vs-All method [27, 35], which reduces multiclass classification to binary classification. Each binary classifier is presented here in the form of an equation of dividing surface. Inductive modeling just allows to find the equation of optimal complexity in n -dimensional space of linguistic variables which we discussed above in Section 2.3.

One can download the trial version of GMDH-S and test it using his/her own data [30]. Universities have the possibility to purchase this product free of charge for teaching purposes.

5. Experiments

5.1 Preprocessing, tuning models and post-processing

For the experiments we used GMDH-S mentioned above. It includes the following possibilities for preprocessing:

- data normalization to a given interval, e.g. [-1.0,1.0] or [0.0, 1.0] ;
- data transformation with various functions such as square root, cubic root or arctg to suppress or to strengthen small and large values;
- balancing classes using copying for small classes.

In the process of modeling a user can do the following:

- to select one of GMDH-based algorithms,
- to limit the total model complexity,
- to assign the form of elements in polynomials,
- to define the external criterion.

Speaking about post-processing we mean both various form of visualization for result presentation and the procedure of ensembling. The latter is averaging a set of the best models selected by GMDH-S. The number of models to be averaged is assigned by a user.

In the experiments we used normalization on [0.0, 1.0] and tested data transformation, balancing and ensembling. The external criterion in model selection was 2-fold cross validation. For tuning algorithms we used recommendations [18]. Taken together we considered 96 options of modeling. Table 12 presents the number of variants for each option.

Table 12. Number of variants for different options

<i>Transform.</i>	<i>Balancing</i>	<i>Ensembling</i>	<i>Algorithm</i>	<i>Complexity</i>	<i>Form</i>
3	2	2	2	2	2

Here: Transformation={without transformation, cubic function, arctg function}; Balancing={without balancing, with balancing}; Ensembling={without ensembling, with ensembling}; Algorithm={classical combinatorial, classical polynomial neural network}; the contents of term ‘Complexity’ depends on the algorithms, it can be the number of members in polynomial or the number of neurons in the model; Form={quasi-lineal, quadratic}

It is necessary to emphasize that we did not tested various document presentations in the experiments although such a presentation essentially affects results (sometimes decisively).

5.2 Building classifiers

The distribution of documents between classes for this category is presented in Tables 7 and 8. Therefore the baselines for A-measure are equal 80% for 2 classes and 63% for 3 classes respectively for the category Restaurants, and 75% for 2 classes and 50% for 3 classes respectively for the category Shopping . The first experiments shows that a) balancing gives worse results than its absence; b) the neural network of GMDH-S works essentially better than the classical combinatorial algorithm. The best options and results are presented in Table 13. Here: NN stands for neural network, yes

(Xm) means option of ensembling with X the best models (X is selected manually), 1000 is the number of neurons.

Table 13. Results of modeling for the categories Restaurants and Shopping

<i>Options</i>	<i>2 classes Restaurants</i>	<i>3 classes Restaurants</i>	<i>2 classes Shopping</i>	<i>3 classes Shopping</i>
Transformation	arctg	arctg	No	cubic
Balancing	no	no	No	no
Ensembling	yes (18m)	yes (8m)	yes (6m)	yes (10m)
Algorithm	NN	NN	NN	NN
Complexity	1000	1000	1000	1000
Form	quadratic	quasi-lineal	Quadratic	quadratic
<i>Results</i>	<i>2 classes Restaurants</i>	<i>3 classes Restaurants</i>	<i>2 classes Shopping</i>	<i>3 classes Shopping</i>
Accuracy	0.91	0.73	0.86	0.71
Defensible accuracy	0.95	0.83	0.97	0.85
Baseline	0.80	0.63	0.75	0.50

The accuracies reached in the experiments are close or exceed those presented in other publications related to opinion mining Yelp review. See, for example [4]. However such a comparison is slightly incorrect because we consider classifications on 2 classes and 3 classes instead of 5 classes. So, to evaluate the results we should take into account other indicators such as defensible accuracy and baseline. With this point of view the results can be assumed as the promising ones, but which should be improved using extended lists of key-terms.

6. Conclusions

In the paper we propose the simple technique for classification of reviews from the well-known Yelp dataset [40]. This technique is based on a) key-term selection using term specificity, and b) construction of predictive models using inductive modeling with GMDH. We consider classifications on 2 classes and 3 classes instead of 5 classes related to 5-star rating. We suppose that in many cases these classifications, namely {unsuccessful, others} and {worst, satisfactory, best}, will prove be more useful for practical applications. We introduce so-called defensible accuracy of results that takes into account the variety of opinions of users with respect to the same materials of social networks. We demonstrate on practical example how to calculate this indicator and we postulate its value as a justified accuracy. We hope this approach will be useful for those who deal with processing data of social networks.

For the experiments we use the well-known tool GMDH Shell and we tune its parameters to build the best model. The results prove to be promising having in view both the simplicity of the proposed technique and the mentioned defensible accuracy.

In the paper we did not test various form of document presentation. We only applied the way we successfully used many times in our projects. We intend to consider this question in future experiments and we will try to maintain the simplicity of the proposed technique.

References

1. Alexandrov, M., et.al.: Inductive Modeling in Subjectivity/Sentiment Analysis (case study: dialog processing); In: Proc. of 3-rd Intern. Workshop on Inductive Modeling (IWIM-2009), Krynica, Poland (2009) 40-43
2. Alexandrov M., Danilova V., Koshulko O, Tejada J.: Models for opinion classification of blogs taken from Peruvian Facebook. In: Proc. of 4-th Intern. Conf. on Inductive Modeling (ICIM-2013), Publ. House NAS of Ukraine & Czech Tech. Univ, Kyev, (2013) 241-246
3. Asghar N.: Yelp dataset challenge: review rating prediction. CS886 Project Report, arXiv: 1605.05362v1 [cs.CL] 17 May 2016, (2016) 9, URL <https://arxiv.org/pdf/1605.05362.pdf>
4. Baeza-Yates R, Ribero-Neto B.: Modern Information Retrieval. Addison Wesley, (1999)
5. Chen P.Y., Wu S.Y., Yoon J.: The impact of online recommendations and consumer feedback on sales. In Proc. of the Intern. Conf. on Inform. Systems, (2003) 711-724
6. Pavlov A., Stepashko V., Kondrashova N.: Effective methods of model self-organization. Publ. House 'Academ Periodica', Kyev, (2014) [rus]
7. Farlow S.J.: Self-Organizing methods in modeling: GMDH type algorithms. Statistics: A series of textbooks and monographs, Book 54, CRC Press, 1-st edition, (1984)
8. Gupta N, Singh S.: Collective factorization for relational data: an evaluation on the Yelp datasets, URL https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_CollectiveFactorization.pdf, (2016) 11
9. Hennig-Thurau T., et al: Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the Internet. Journ. of Interactive Marketing, vol.18, No.13, (2004) 39-52
10. Huang J., Rogers S., Joo E.: Improving restaurants by extracting subtopics from Yelp reviews. In: CS294-1 Spring 2013: final project 1, URL https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_ImprovingRestaurants.pdf, (2013) 5
11. Ivakhnenko, A.: Heuristic self-organization in problems of automatic control. Automatica (IFAC), No. 3, (1970) 207-219
12. Ivakhnenko, A.: Polynomial theory of complex systems. IEEE Trans. System, Man and Cybernetics, No.1(4), (1971) 364-378
13. Karakaya F., Barnes N. G.: Impact of online reviews of customer care experience on brand or company selection. Journ. of Consumer Marketing, USA, URL: <https://www.researchgate.net/publication/235316561>, (2010) 447-457
14. Kaurova O., Alexandrov M., Ponomareva N.: The study of sentiment word granularity for opinion analysis (a comparison with Maite Taboada works). Intern. Journal on Social Media MMM: Monitoring, Measurement, and Mining, Brno, Publ. House 'Konvoj', No.1, (2010) 45-57
15. Kaurova, O., Alexandrov, M., Koshulko, A.: Constructing classifiers of medical records presented in free text form. In: Proc. of 4-th Intern. Conf. on Inductive Modeling (ICIM-2013), Publ. House NAS of Ukraine & Czech Tech. Univ., URL: <http://mgua.irtc.org.ua/attach/ICIM-IWIM/2013/4.8%20.pdf>, (2013) 273-278
16. Kilgariff A.: BNC database and word frequency lists, URL: <http://www.kilgariff.co.uk/bnc-readme.html>
17. Koshulko O., Alexandrov M., Danilova V.: Forecasting Euro/Dollar rate with Forex News. In: Proc. of the 19th Intern. Conf. on Application of Natural Lang. to Inform. Systems (NLDB-2014), Springer, LNCS, (2014) 148-153
18. Koshulko O., Koshulko G.: Validation Strategy Selection in Combinatorial and Multilayered Iterative GMDH Algorithms. In: Proc. of 4th Intern. Workshop on Inductive Modeling (IWIM-2011), NAS of Ukraine, Prague Tech. University, Kyev, URL: <http://mgua.irtc.org.ua/attach/ICIM-IWIM/2011/7%20.pdf>, (2011) 51-54

19. Leung C.W.K., Chan S.C.F.: Sentiment analysis of product reviews, In: Report Hong Kong Univ, URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.1549&rep=rep1&type=pdf>, (2006) 15
20. Linshi J.: Personalizing Yelp star ratings: a semantic topic modeling approach. In: Report of Yale University, URL https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_PersonalizingRatings.pdf, (2014), 9
21. Liu J., et al: Mining quality phrases from massive text corpora. In: Proc. SIGMOD-15, URL https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_MiningQualityPhrases.pdf, (2015) 16
22. Lopez, R., et.al.: Lexistern – the program for term selection by the criterion of specificity. In: Artificial Intell. Application to Business and Engineering Domain, ITHEA Publ., Rzeszov-Sofia, vol.24, (2011) 8-15
23. Manning C., Raghavan P., Schütze H.: Introduction to Information Retrieval. Cambridge Univ. Press, (2008)
24. Makagonov P., Alexandrov M., Sboyachkov K.: A toolkit for development of the domain-oriented dictionaries for structuring document flows. In: Data Analysis, Classification and Related Methods, Springer, series Studies in Classification, Data Analysis, and Knowledge Organization, URL https://link.springer.com/chapter/10.1007/978-3-642-59789-3_13, (2000) 83-88
25. McAuley J., Leskovec J: Hidden factors and hidden topics: understanding rating dimensions with review text, In: Proc. of the 7-th ACM Conf. on Recomm. Systems, URL <http://infolab.stanford.edu/~julian/pdfs/recsys13.pdf>, (2013) 8
26. Madala H., Ivakhnenko A.: Inductive learning algorithms for complex systems modelling. CRC Press, (1994)
27. One-vs-All, Wikipedia, URL: http://en.wikipedia.org/wiki/Multiclass_classification
28. Pang B., Lee L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proc. Conf. on Empirical Methods in Natural Lang Proc., (2002) 79-86
29. Pang B., Lee L.: Opinion mining and sentiment analysis, In: Foundations and Trends in Information Retrieval, vol. 2, No 1-2, (2008) 1-135
30. Platform GMDH Shell: URL <http://gmdhshell.com>
31. Resource GMDH: Dept of Inform. Tech. in Induct. Model. NAS of Ukraine, URL <http://mgua.irtc.org.ua/>
32. Stein B., Eissen S.M., Wibbrock F.: On cluster validity and Information needs of users. In: Proc. 3-rd IASTED Conf/ on Artif. Intell. And Appl. (AIA-2003), Acta Press, (2003) 216-221
33. Stepashko, V.: Developments and prospects of GMDH-based Inductive Modeling. In: Proc. of 12-th Intern. Scientific and Technical Conf. on Computer Science and Inform. Technologies (CSIT-2017), Springer, LNCS, (2017) 18
34. Stepashko, V.: Method of critical variances as an analytical tool of the Inductive Modeling theory. Journ. of Inform. and Automat. Sciences, Begell House Inc, vol. 40, N_3, (2008) 47-58
35. Tax D.M.J., Duin R.P.V.: Using two-class classifiers for multiclass classification. In: Proc. of Intern. Conf. on Pattern Recognition (Quebec, Canada), IEEE, URL <http://prlab.tudelft.nl/content/using-two-class-classifiers-multiclass-classification>, (2002) 1051-1054
36. Tweedie F.J., Baayen R.H.: How Variable May a Constant be? Measures of Lexical Richness in Perspective. In: Computers and the Humanities, Kluwer Academic Publishers, vol. 32, (1998) 323–352

37. Taboada M., Anthony C., Voll K.: Creating semantic orientation dictionaries. In: Proc. of 5th Intern. Conf. on Language Resources and Evaluation (LREC), Italy, (2006) 427-432
38. Turney P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of 40th Meeting of the ACL, (2002) 417-424.
39. Walker E., Brown. A.: What success factors are important to small business owners. In: Intern. Small Business Journal. vol.22, No.6, (2004) 577-594
40. Yelp Dataset: URL http://www.yelp.com/dataset_challenge, (2014)