Theory of Semantic Field for Sentiment-Analysis of the Language of Specific Users' Group in Social Media (Case of Freelancer Groups)

Anna Maltseva, Natalia Shilkina, Igor Tiomniy
Saint-Petersburg University
Saint Petersburg, Russia
a.maltseva, n.shilkina, i.tiomniy @ spbu.ru

Olesia Makhnytkina, Inna Lizunova
ITMO University
Saint Petersburg, Russia
makhnytkina, lizunova@itmo.ru

***Abstract*—In this article the description of an algorithm of a statement sentiment evaluation is done for the users' of social media language. We underline that statements of the natural language can be contradictory, emotionally complicated, ambiguous. It is the additional research task to detect the adequate formal criterion of the natural language statements ranging on the scale "positive – negative". In the article the original decision of this problem on the base of the theory of the semantic field is described. The technique was tested in the All-Russian Scientific Research Institute of Labor of the Ministry of Labor and Social Protection of the Russian Federation for investigation of population opinion to the new forms of employment. Empirical base is more than 100 000 messages of users of thematical groups in VKontakte. Analysis with the accent on the parameters: subject of tonality, object of tonality, message tonality was done. The technique of research assumes the detection of words-markers that indicate the general message tonality.**

I. Introduction

Social media sites are a source of data that brightly reflect the social reality and its main tendencies from the users' language perspective. Such data let us to analyze users' lexicon from the point of its semantic specifics, to build the semantic field and to make the special linguacultural comment for the language units that in fact are the basic concept of the social media lexicon.

The most actual perspective of the social media lexicon analysis is the sentiment-analysis that aims to detect the subjective opinions all along with the tonality or opinion orientation of a forum participants statements.

Research problem is quite simple if we plan to analyze social media sites for marketing tasks. For this goal we can compare any users' statements about any goods or service with the special formal symbols – such as grades/ marks, asterix, smiles or emojis etc., so it makes the process of the statement's emotional features evaluation (sentiment-analysis) easier. But if the research problem is in the evaluation of a statements of a discussion on social or political topic that touch one's personal values or cultural semantic components it is not easy to find obvious and unambiguous formal indicators of tonality differentiation. Also, such sort of statements is consisted of more complicated and contradictory sub-parts, with allegories and ambiguities. In real communication the language can reflect the wide emotional shades such as irony, mockery, skepticism, mistrust, hope and infinite number of others. But the virtual on-line communication is much more difficult and unusual because of the specific ways of the context all along with the sub-contexts such as world outlook, political, social and ideological besides all these components have special substantial attributes: temporary duration, symbolic, message volume.

Researcher have to outline an adequate formal criterion for the further range of each statement on the scale "positive – negative". Validation of the procedure of such evaluation is the next problem for the high-quality sentiment-analysis in the manner of impartial assessment. Finally, each statement should be evaluated as positive, neutral or negative, also it is possible to range the statements more thorough – differ as less negative or less positive. In this article we describe our decision of the problem of the natural language formalization and outline the most important from our point of view nuances of the natural language for detecting of statements tonality.

II. Related works

Researchers apply number of methods for sentiment-analysis of users' messages in social media and choosing of the approach depends on the research goals, the uniqueness of the analyzing data semantics, availability of marked data sets on the topic. As usual researchers get data from the Tweets [1]-[3] when the length of the message is very important (since 2017th it is possible to make Tweets till the 280 signs), at the same time the popular in Russia social media VKontakte it is possible to write messages till the 15895 signs for post "on the wall" and 4096 signs in the section "my messages" [4], [5] also there are some specific for the messages from the hidden social media [6].

Most general classification of approaches to the sentiment-analysis divides all of them to the following groups: a) based on rules, b) based on dictionaries, c) methods of machine learning (supervised and unsupervised learning), and d) hybrid method. Methods of machine learning is getting more popular. Among methods of machine learning can be mentioned Bayesian networks [7], an LSTM (Long-Short Term Memory)

network sentiment polarity analysis method [8], [9]; fuzzy C-means algorithm based on Simulated Annealing (SA-FCM) to cluster the explicit comment sentences into classes [10]; unsupervised learning methods [11]. Application of some methods requires using of vector form for words, texts [1]; among them are pre-trained vector models of words Word2Vec и GloVe, that can be used as a base for the original models of public opinion analysis [12].

Very important stage in development of machine learning methods for sentiment-analysis is applying of deep learning approach [13]. For example, tensor-based, tree-structured deep neural network (named Discourse-LSTM) in order to process the complete discourse tree [14].

The approaches based on the rules is classical for solving the very specific task and the data set is not large [15], [16]. Important role here plays the special vocabularies\ thesaurus - the base for the rule constructing [17], [18]. Benefits of the machine learning and rule-based approaches can be combined in hybrid approach [19].

For properly understanding of modern social phenomenon, analyzing public opinion through the additional data sources and from the point of view of different specific group of social media users it is very important to investigate language specifics of such groups and form adequate tools of evaluation of their opinion orientations [20]-[23].

Correct evaluation of opinion orientation suggests detecting of classification point. Definitions of such grounds depends on researchers' scientific interests or/and the existing practice. Reviewing of the scientific publications for the last years demonstrates us the actuality of the searching for decisions of the sentiment-analysis of complicated sentences [24]-[28].

In this article we suggest application of the theory of the semantic field that let us to evaluate the opinion orientation (or sentiment) in a manner of "point estimate", to eliminate the non-relevant messages and to differ relevant ones to "the core" and "the periphery" messages. As usual for the sentiment-analysis of the natural language it is necessary to solve two interim targets: 1) making of the unique vocabularies for each specific case; 2) an increasing of the range of using rules. We considered that text of a message can include the individ (subjective) opinion on it directly ("the core"), but also the text of the message about the object of tonality can include the individ (subject's) opinion just formally ant its expressive/ emotional meaning will relate to some other semantic objects ("the periphery"). This perspective allows us not just to evaluate the message tonality but make the pointed (direct) evaluation of the user's attitudes to the object of the tonality. And considering the place of the object of the tonality in the original message allows us to do the evaluation impartial.

III. THEORETICAL PERSPECTIVE AND MAIN TERMS OF THE RESEARCH

Identification of the opinion orientation of a message is compounded by the existing philosophical, political, ideological sights of an individ or can be impartial, direct, assessing, analyzing [29],[30]. Evaluative positive tonality (allegation, claim, conviction, opportunity, certainty etc.) according to the semantic of the analyzing units is close to the meaning of word "good". Neutral tonality – an informative message or factual particulars usually presented through the language units without any emotional expressive coloring. Evaluative negative tonality according to the semantics of analyzing units is close to the word "poor\ bad" or the same [31], [32], [33]. Tonality of a text can be determined through the tonality of "core" words and the words around.

Author approach is based on the theory of the semantic field and therefor the system of interactions of the sub-fields allows one to reveal "core" part of the lexicons. The presence of this "core" part is the marker that define the correlation of this text to the certain tonality.

Tonality is the emotional expressive color (meaning) of a message. On the base of the theory of the semantic field we will list below the parameters of tonality for the group of freelancers.

Object of tonality – lexical units that are used in groups of freelancers in social media.

Tonality of a message – measure of an expressiveness of the message on the scale "positive – negative".

Lexical markers - morpheme (minimally meaningful unit of language allocated from a word) that defines the tonality of the statement.

Semantic field – association of lexical units according their general topic "freelance".

Semantic sub-fields – thematical parts of the semantic field "freelance".

One more parameter – "a subject of a tonality" is interesting for the analysis as an actor that verbally expresses his/her attitude to an object (process). That is why the meanings of the lexical unit "work" are different depends of the subjects of the tonality. For example, among the groups "Overheard in …" with the discussions of students problems the lexeme "work" in the meanings of "test/ supervised tasks", "occasional paper", "senior thesis", "paper/ abstract" [34]; in the groups of film fans the range of meanings for the lexeme "work" includes – "film", "activity of the film crew"; in the groups of freelancers – "order", "vacancy", "labor activity". These differences demonstrate us the importance of making specific vocabularies that relevant to the subject of the tonality. At the same time for the accurate analyzing of the opinion orientation we must pay the attention to the subject of the tonality as the author not in the context of his/ her social or demographic features but as a resident of the interesting group. Because, for example, the problem of freelance can be discussed by freelancers and other "interested faces".

Analyzing of the tonality subject's messages is carried out in the light of the recurring lexical expressions; fragmented and whole speech constructions; syntaxis of appeal.

IV. Empirical base of the research

A. Datasets

Collecting of the data was organized with the authors original WindowsForm software that can function with the API interface of VKontakte that allows to get data from the VK-data base with the http-requests to the special server. Empirical data include text of messages with the all related characteristics (likes, reposts etc.).

Empirical base includes more than 100 000 messages of users form the specialized group “freelancer” and users that express their opinions about this new form of employment. All messages cover one calendar year. The number of the messages was limited only by the users’ activity in discussions about “freelance”. As the main groups were considered "Freelancers Club"; "Entrepreneur. Community of successful and aspiring entrepreneurs!"; "Guild of media freelancers". The group "Guild of media freelancers" is a professional Association of media representatives of St. Petersburg, who are not in the editorial staff, television or radio broadcasting company, publishing house. The Guild includes members of the Union of Journalists of St. Petersburg and Leningrad region and colleagues who share the principle of remote cooperation.

The group "Freelancers Club" was created to help freelancers and everyone who wants to try their hand at freelancing, share experiences, seek customers and improve their professionalism, maintain active communication, discussions and exchange of experience, if they are held in a civilized and respectful atmosphere.

In the group "Community of successful and aspiring entrepreneurs!" useful, educational, motivating and necessary information for conducting and promoting various types of business is published. The purpose of this business community is to bring together all entrepreneurs to exchange experience and knowledge.

At the first stage of the data processing for preliminary identification and analysis of possible topics in message texts containing users' opinions on freelance, the Dirichlet latent placement method (LDA) was used. LDA belongs to the family of generating probabilistic models in which topics are represented by the probability of occurrence of each word from a given set. In total, 5 main topics were identified that reflect freelance as a form of employment (class 1), personal life of a freelancer (class 2), vacancies and job offer (class3), work modes (class 4) and freelance as a way of earning (class 5) (Table I).

TABLE I. LDA model

Topics	Key words
Topic 1	'0.015*" (this)" + 0.014*"freelance" + 0.011*"(which of)" + 0.011*" (work)" + 0.010*" (own)" + 0.008*" (to work)" + 0.006*"(person)" + 0.006*" (cash)" + 0.006*"(time)" + 0.005*"(one)"')]
Topic 2	'0.016*"(this)" + 0.013*" (life)" + 0.011*" (person) " + 0.009*" (child)" + 0.008*" (year)" + 0.007*" (which of)" + 0.006*" (work)" + 0.006*" (to live)" + 0.005*" (can do)" + 0.004*" (time)").
Topic 3	'0.070*" (work)" + 0.010*" (vacancy)" + 0.009*" (Moskow)" + 0.009*" (year)" + 0.009*" (to search)" + 0.006*"Internet" + 0.006*" (work up)" + 0.006*"(home)" + 0.006*" (our)" + 0.005*"resume"')
Topic 4	'0.011*"online" + 0.011*" (this)" + 0.010*" (day)" + 0.006*" (morning)" + 0.006*" (work)" + 0.006*"(time)" + 0.004*"(own)" + 0.003*" (very)" + 0.003*"(volume)" + 0.003*" (all)"').
Topic 5	'0.027*"rouble" + 0.021*" (cash)" + 0.020*" (own)" + 0.018*" (year)" + 0.016*" (this)" + 0.012*" (to invest)" + 0.011*"(work)" + 0.010*"(month)" + 0.009*"(annual)" + 0.008*"(distant)"').

The map of the distances between the topics shows that topics 1 and 2 are quite close and include the largest number of posts, the remaining topics contain a significantly smaller number of posts and are quite distant from each other. (Fig. 1).

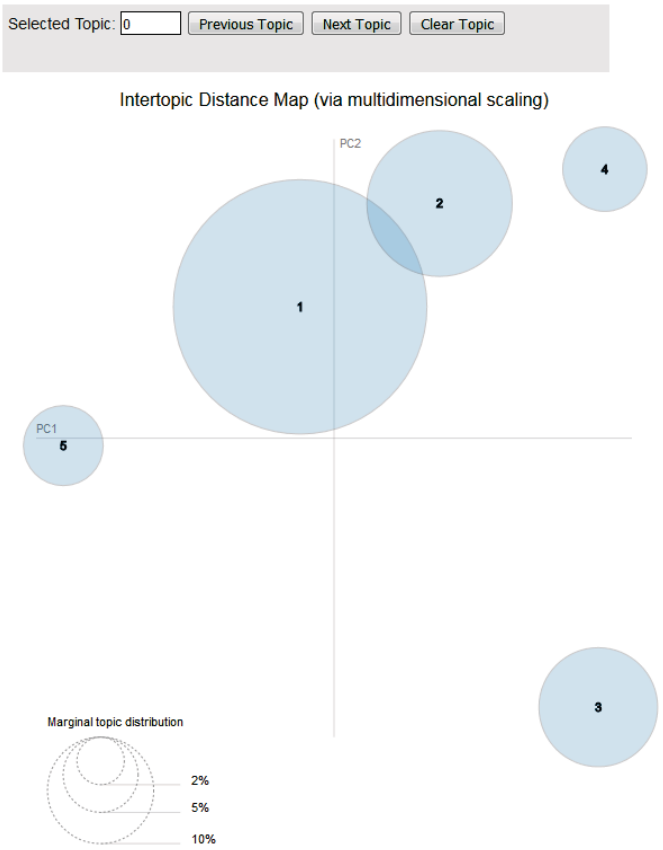

Fig. 1. Topic of the messages

Topic 1. Here the word "freelance" is thematically tied to nouns: "work", "money", "time" and the verb: "work", which shows the identification of freelancing activity with traditional forms of employment, with labor, generating income. Time expenses are shown: “time” and a certain responsibility is indirectly shown: “own”, “one”. The word “man” in this connection apparently fulfills the function of the word “worker” as in Russian classical literature of the 19th century, when the appeal “man” was accepted for workers impersonally fulfilling the task of clients, for example, ordinary employees of a tavern or an inn.

Topic 2. The second thematic group of words reveals the saturation of life outside the work of a freelancer. These are the words “life”, “child”, “year”, “be able” “time”. The word "time" is included in the topic of freelance discussion as a form of employment and as life outside of work. The acute

problem of freelancer time distribution and self-organization is likely, as well as the problem of interacting with children when "you are at home, but you are at work." In general, the problem of time with the transition from traditional employment to freelance does not disappear, but changes and requires the formation of new behavioral practices.

Topic 3. The third thematic group of words is formed in the discussion of vacancies. This set seems to paint a picture of the search for “vacancies” for “work” and “part-time” from “home” by sending a “resume” via the Internet. It is logical that at present the search is primarily in Moscow as a business and financial center with a high rental cost of office space.

Topic 4. The word “time” again falls into the selection of words that reflect this time the “online” operating modes. At the time of day, “morning”, “day” is mentioned, which reflects the identity of work online and offline.

Topic 5. This thematic group reflects the characteristics of freelance as a way of remote earnings, combining the words: “work”, “ruble”, “money”. The names of money that are used here are somewhat dismissive in nature, which is achieved by distorting the vocabulary form of the word “money” as “money” and using the ruble to denote the smallest unit. This is possible with a low level of income and difficulties in finding income, factors that impede the life of a freelancer, which he hides behind a neglect of earnings as a result of his work, focusing on its content and original, unconventional registration of labor activity. It also discusses the calculation of earnings - as "annual" or in the "month" as is traditionally accepted.

At the second stage of data processing messages from the above-mentioned groups of social media were reviewed by experts and their tonality was assessed on the basis of the formulated set of rules (Table II).

TABLE II. STATISTICS OF MESSAGES’ TONALITY

Group	Number of messages	Number of positive messages	Number of neutral messages	Number of negative messages
“Freelancer club. Freelance is a cool!”	4610	1040	2292	1278
Community of successful and aspiring entrepreneurs	4132	1663	1567	901
"Guild of media freelancers"	231	97	93	41

The rest of the messages from the total sample of data contained the opinion about freelancing as a form of employment.

B. The research technique

On the base of the theory of the semantic field we suggest following research technique. Each lexical unit from the message is included in one of three possible groups: “lexical dominant core”, “peripheral lexical units”, “irrelevant lexical units”.

A lexical unit can be inserted into the dominant core if it contains the following indicators: 1) maximal concentration of the specific marks (tokens) – e.g. being the core in the professional group or given units in the context of the analysis goal; 2) possibility of the maximal opposition on the scale “core – periphery”; 3) maximal functional loading in the core along with the weakening in the periphery; 4) high level of the semantic function implementation; 5) regularity of the definite language marks, high frequency on its using.

A lexical unit can be inserted into the periphery if it has only semantic tie with the lexical unit from the dominant core.

A lexical unit inserted into the irrelevant if it does not present with the identical meaning frequently. For example, the researcher would like to analyze the group of freelancers in one of the social media and should to mark as irrelevant all lexical units that does not have direct relation to the freelance. Such kind of messages can inform about its author’s plans to get married or any other things of that (without any connection with the topic of the analysis.

Compare the core dominant on the base of their relationship with the extralanguage reality e.g. with the things they mean or call, researcher can find out the dual nature of word meaning: any word at the same can be a sign of the reality and a unit of language, it defines a thing out language and relates with different parts of language. It is also important to note that the type of the lexical meaning: direct or nominative; connected by a phrase or phraseologically related; syntactically caused [35]. Ties between direct\ nominative words are based on the subject-logical relations and not on the lexical. This fact demonstrates the unlimited possibility for ties between lexical units with the direct or nominative meanings.

Consequently, if a word presents in the same mean in all cases of it using in the message with the accent on the defined context than it can be inserted into the core of the field. If there are some new colors of the lexical units meaning inside the analyzing semantic field, then we mark such units as irrelevant ones. All other words with the related meanings we mark as peripherical ones. The list of core units has the enshrined tonality meanings. They are lexical markers that influence on the tonality evaluation of the message in general.

Detection of the lexical units makes from the users’ messages analyzing. Also the following metadata are collected: 1) “time of living” of the theme; 2) importance ot the theme during its “life time”; 3) presence in the message of any specific signs of emotional expression – e.g. smiles, emojies, CapsLock using, underling or strikethrough etc.; 4) presence in the message of unigramms as the most important verbal marker of the lexical unit meaning; 5) time of the message publishing on the site of a social media; 6) presence of several unigrams in the theme that are combined in bigrams and triplets.

All lexical units have been submitted in the form of unigrams, bigrams and triplets. Unigrams were defined on the base of expert analysis as the most important verbal markers for the analyzing users’ group. Bigrams contain all possible pairs of unigrams with the one that was defined by the expert analysis as the most important verbal marker. Triplets contain

all possible combinations of the unigrams when one of unigram was defined by the expert analysis as the most important verbal marker.

C. Definitions of the sub-fields

The decision of the sub-field choosing for the sentiment-analysis depends on the research goal. For example, if the research is about the freelance as a form of employment then as the base for the core of the field meanings presenting the official document should be chosen. So, we took the Russian Labor Code. On the base of the chosen document or any other ground defined by the researcher the body of the lexical markers are formed. Below we present short fragment of such list of lexical markers made on the base of the Russian Labor Code [36] for sentiment-analysis of the freelance theme (Table III). Lexical markers listed in Tab. 1 were got with the expert opinion of specialists in employment and labor sphere, as well as linguists.

TABLE III. FRAGMENT OF THE LEXICAL MARKERS LIST

Sub-fields	Lexical markers	Samples of the original messages with lexical markers
Fair wage	Money	Money in three months? Result in three months. :)
	Earnings	Freelance – is the sort strange practice when you some earn more during just one evening after work when at the work…
	Salary	Friends, just found in the Internet: "40% of the staff members will agree with retrenchments if they will offer telecommuting and they will be free from everyday visiting of office…"
	Income	Amounts less than 100 000 rubles is not good income
	Retainer	Your client does not pay the retainer? Deprive him/ her the driver license. New draft law proposes to deprive the driver license of those who does not pay alimony, salary or retainer
Contractual agreements in labour relations	Work	Wants a photographer! Special work! Friends, for one or our project we looking for an architectural photographer. Please, write to e-mail. Portfolio is necessary. Looking for somebody who able to do like this: http://www.ratusha.ru/gallery/
	Vacancy	#vacancy Wanted a journalist-author for a sturt-up blog. Freelance.
	Order	Sites of services for freelancers. Would you like to work with a simple order at a fixed price?

Expert analysis allows to find out actual for users' words-topic or subfields all along with the problems that users of different Internet- "community" care about (in our case – freelancers). In this way we can define core units of the users' picture of the world. This core units are strongly significant for a single user and for the semantic field in general.

In addition to the dominant-markers that were got by the expert opinion it is important to analyze the dominant-markers that we can get frequencies of their appearance. The stage of frequency analysis consists of the following steps: 1) defining of group of dominant unigrams (according the frequencies of their appearance), threshold values for them we can define with significant for the target users' group periods of time (e.g. news, bills, events, rumor etc.); 2) detecting of the semantic components of unigrams – detailed study of unigrams changing, development specific and the current condition; 3) matching the data on the different sub-field according to the each group of unigrams.

Analyzing of the three specified social media users' group "Freelancer club. Freelance is a cool!", "Businessman" and "Guild of media-freelancers" (VKontakte) the experts' dominant-markers list (vocabular) of bigrams and triplets has been expanded by the most frequent ones (we present fragment of this list in Table IV).

TABLE IV. FRAGMENT OF THE MOST FREQUENT LEXEMES/N-GRAMS

Bigrams	Count	Triplets	Count
Group "Freelancer club. Freelance is a cool!"			
'vacancy', 'web-site'	604	'test', 'reliability', 'customer'	471
'add', 'vacancy'	510	'reliability', 'customer', 'help'	471
'reliability', 'customer'	477	'customer', 'help', 'tool'	471
'amount', 'work'	475	'large', 'volume', 'work'	470
'take', 'prepayment'	473	'fulfill', 'large', 'volume'	466
'large', 'volume'	473	'volume', 'work', 'preliminary'	466
'test', 'reliability'	471	'work', 'preliminary', 'payment'	466
'customer', 'help'	471	'add', 'vacancy', 'web-site'	451
'help', 'tool'	471	'take', 'prepayment', 'negotiate'	432
' fulfill ', 'large'	466	'prepayment', 'negotiate', 'incremental'	432
Group "Businessman"			
'successful', 'man'	67	'move', 'our', 'partnership'	14
'your', 'company'	58	'our', 'partnership', 'link'	14
		'change', 'your', 'life'	10
		'equation', 'creating', 'title'	10
		'your', 'good', 'service'	10
Group "Guild of media-freelancers"			
'union', 'journalist'	62	'union', 'journalist', 'Saint-Petersburg'	28
'guild', 'media-freelancer'	51	'meeting', 'guild', 'media-freelancer'	21
'meeting', 'guild'	25	'guild', 'media-freelancer', 'union'	20
'media-freelancer', 'union'	20	'media-freelancer', 'union', 'journalist'	20

V. SAMPLES

A. Samples of the freelancers' group slang lexical units

As a part of special lexica of the freelancers group in the above mentioned social media sites can be detected the branch-specific systems of terms: "freelance", "business", "development", "independence" with the organizational relations between the terms based not on the language principals but on the subject and logical ties between the related concepts that finally reflect the structure of the analyzing subject. For this special lexicon is typical the higher level of the inside system organization of its parts e.g. the branch-system due to the existence of the related concepts. Also quite usual the presence of abbreviate forms of unigrams ('zp'/ salary, 'comp'/ personal computer etc.); abbreviations ('CZN'/ employment center, 'IP'/ individual proprietor etc.);

professional slang (‘codit’/ writing a program code, *‘tizhecopywriter’/ you are copy writer etc.);* homely phrases (*‘shalai-valai’/ doing somehow, ‘zashibit dengu v Seti’/ hurt the many in the Network etc.)*; borrowed forms of English ‘*(dedlain’/ deadline, ‘brif’/ breefing, ‘laik’/ like etc.);* lexemes with the multiple semantic changes with the unchanged final form (‘sharit’/ changed form of English word ‘share’ means searching for, ‘upakovannii’/ changed form of English word ‘packed’ means “finished well or looking well” etc; indefinite pronounce *(‘kuda-to’/ somewhere, ‘chto-to’/ something, ‘ kto-to’/ somebody etc.)* [37].

B. Samples of the detecting of the statement tonality

Here we would like to demonstrate an example of the message tonality. “*There is a simple way for people – freelance in any demanded sphere. An unemployment us because the (people) want (to have) much but are able to do very few. If it would be so bad in Tambov than there would not be so much cars (on the city streets). Having salary about 10 thousand rubles is not enough even for buying and maintain an old car like Zhiguli”.* Lexical unit “unemployment” here on its lexical meaning matters negative coloring. Lexical unit “a few” matters negative coloring on its context meaning. In lexical unit “not enough” negative color comes from negative particle. Consequently, in amount of all word-markers negative tonality of this message can be stated.

VI. Conclusions

By setting the definite frame from the basic points – dominant words (freelance, tax, pension, reform, income etc.) we analyzed messages considering the texts’ semantical, lexical, morphological and syntax specifics and in accordance with the context of each message detected its tonality.

One of the research preliminary score is the detection of tonality of several lexemes. Lexeme “freelance” in the most contexts detected in the meaning “development”, “business”, “independence”, “do the things I like” and only than in a mean “earning”. Very often the concept “freelance” is in opposition to word “work” with negative coloring (“His job is like any other, but my work is freelance”). In number of the messages detected negative coloring due to the freelance evaluating by the third parties (like in discussion of bank officer with a freelancer: *“- What is your occupation? – I’m a freelancer. - And what’s that?"*»).

Lexeme “tax” is in the contexts with negative tonality, lexemes “reform” and “changes in legislation” in neutral and positive. Frequency of lexeme “pension” using in the freelancers’ group detected in active field with neutral coloring. All above mentioned let us to conclude that participants of the freelancer’s forums point out the necessity of changes in the legislation but demonstrate negative opinion about the idea of tax for freelancers and the idea of pension provision is not significant for the freelancers.

Lexeme “self-employment” was detected in the positive context as a way to increase the income, to plan the working schedule etc. There are lots of contexts with successful results in personal and financial life-stories of different people. But still this context full of indefinite pronouns “she writes for some start-up blog; he reads some lectures somewhere etc.”.

Lexeme demonstrates neutral tonality in the most messages. At the same time the lexeme “work” in bigrams “work for the man” has negative connotation and highly frequent.

The described technic of the social media contents seems to be a useful tool for the analyzing of stereotypes, expectations and potential claims of different groups of the on-line communities, as well as the differentiations in opinions and system of values, understanding and evaluating messages properly.

In the perspective the research will be continued in two main directions: 1) development of the functionality of the research approach and the original soft-ware; 2) application the approach to the biggest data sets from the other social media specified group of users with the goal to collect more core parts of their lexicon.

Analyzing the content of the specified users’ group in social media will help to detect specifics of the semantic fields of their language. It will provide the solving of the task of dominant-markers vocabularies (including hierarchical vocabularies) widening through the collecting of the lexical units from the messages. Collecting of such vocabularies will improve the quality of the opinion orientations evaluating in the context of the messages’ object. In perspective it will be possible to find an instrument for description of language picture of the world for different types of users. Also, these results, based on the lexemes tonality would be useful for diagnostics of the level of social tension and potential readiness of people to the social changes

For developing of the functionality of the approach we plan to widen the scale for evaluating of the message tonality from three (negative, neutral, positive) to seven (strong positive, medium positive, weak positive, neutral etc.). Also, it will help to count an integral mark of users’ opinion towards any object or process that will consider not only the message tonality but also its evaluation by other users (likes, dislikes, number of comments, reposts, time of the message “life”). Such indicator seems to be useful for the impartial monitoring of the changes in public opinion towards the object of tonality in the messages besides it can help to analyze the opinion orientation of those users that do not create any original messages but only repost or comment others.

Finally, collecting of big amount of the marked data sets will allow us to use combined approach to the sentiment-analysis of the opinion towards the objects of the tonality on the base of the machine learning methods.

Acknowledgment

This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.575.21.0178 (ID RFMEFI57518X0178).

REFERENCES

[1] S. K. Biswa, "Keyword extraction from tweets using weighted graph", *Cognitive Informatics and Soft Computing,* January 2019*,* pp. 475-483.

[2] P. Pugsee, V. Nussiri, W. Kittirungruang, "Opinion mining for skin care products on twitter", *Communications in Computer and Information Science* vol. 937, 2019, pp. 261-271.

[3] N. Ravishankar, R. Shriram, "Grammar rule-based sentiment categorization model for classification of Tamil tweets", *International Journal of Intelligent Systems Technologies and Applications* vol. 17(1-2), 2018, pp.89-97.

[4] R. V. Posevkin, I. A. Bessmertny, "Texts sentiment-analysis application for public opinion assessment", *Scientific and technical journal of information technologies, mechanics and optics*, vol.15(1), 2015, pp. 169–171.

[5] A. Rogers, A. Romanov, S. Rumshisky, M. Volkova, A. Gronas, A. Gribov, "RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian" in: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, 2018, pp. 755–763.

[6] R.M. Alguliyev, R.M. Aliguliyev, Niftaliyeva G.Y., "Extracting social networks from e-government by sentiment analysis of users' comments", *Electronic Government* vol.15(1), 2019, pp.91-106.

[7] L. Gutiérrez, J. Bekios-Calfa, B. Keith, "A review on bayesian networks for sentiment analysis", *Advances in Intelligent Systems and Computing*, vol.865, 2019, pp.111-120.

[8] Z. Chen, S. Teng, W. Zhang, H. Tang, Z. Zhang, J. He, X. Fang, L. Fei, "LSTM sentiment polarity analysis based on LDA clustering", *Communications in Computer and Information Science* vol. 917, 2019, pp.342-355.

[9] H. Kaya, D. Fedotov, A. Yeşilkanat, O. Verkholyak, Yang Zhang, A. Karpov, "LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition", in *Proc. Interspeech*, Hyderabad, 2018, pp.521-525.

[10] Z. Liu, Q. Shen, J. Ma, "Extracting implicit features based on association rules", in *Proceedings of the 3rd International Conference on Crowd Science and Engineering*, New York, 2018, pp. 1-7.

[11] S. Popova, I. Khodyrev, I. Ponomareva, T. Krivosheeva, "Automatic Speech Recognition Texts Clustering", in *Text, Speech and Dialogue. Lecture Notes in Computer Science*, Cham, 2014, pp. 489-498.

[12] S.M. Rezaeinia, R. Rahmani, A. Ghodsi, H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings", *Expert Systems with Applications* vol.117, 2019, pp. 139-147.

[13] H.H. Do, P.W.C. Prasad, A. Maag, A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review", *Expert Systems with Applications* vol.118, 2019, pp.272-299.

[14] M. Kraus, S. Feuerriegel, "Sentiment analysis based on rhetorical structure theory: vLearning deep neural networks from discourse trees", *Expert Systems with Applications* vol. 118, 2019, pp. 65-79.

[15] S. Rani, P. Kumar "Rule based sentiment analysis system for analyzing tweets", in *International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions (ICTUS 2017)*, Dubai, 2018, pp. 503-507.

[16] C. Wu, D. Zhang "Ranking products with IF-based sentiment word framework and TODIM method", *Kybernetes*, vol. 48 No. 5, 2019, pp. 990-1010.

[17] A. Dey, M. Jenamani, J.J. Thakkar "Senti-N-Gram: An n-gram lexicon for sentiment analysis" *Expert Systems with Applications*, vol.103, 2018, pp. 92-105.

[18] S. Wu, F. Wu, Y. Chang, C. Wu, Y. Huang, "Automatic construction of target-specific sentiment lexicon", *Expert Systems with Applications* vol.116, 2019, pp. 285-298.

[19] C. Wu, F. Wu, S. Wu, Z. Yuan, Y. Huang. "A hybrid unsupervised method for aspect term and opinion target extraction", *Knowledge-Based Systems*, vol. 148, 2018, pp. 66-73.

[20] G. D'Agostino, F. D'Antonio, A. De Nicola, S. Tucci, "Interests diffusion in social networks", *Physica A: Statistical Mechanics and its Applications*, vol. 436, 2015, pp. 443-461.

[21] P. Peverini "Storytelling and «virality». Marketing communication from a semiotic perspective", *Lexia*, vol.25-26, 2016, pp. 417-439.

[22] C.A. Porcino, M.T.Á.D. Coelho, J.F. De Oliveira, "Social representations of university students on travesti people", *Saude e Sociedade*, vpl.27 (2), 2018, pp. 481-494.

[23] X. Zou, J. Yang, J. Zhang "Microblog sentiment analysis using social and topic context", *PLoS ONE*, vol. 13 (2): e0191163, 2018, pp.1-24.

[24] A. Chatterjee, U. Gupta, M.K. Chinnakotla, R. Srikanth, M. Galley, P. Agrawal, "Understanding Emotions in Text Using Deep Learning and Big Data", *Computers in Human Behavior*, vol. 93, 2019, pp. 309-317.

[25] N.V. Korytnikova, "Online Big Data as a source of analytic information in online research". *Sociological Studies,* vol.8, 2015, pp.14-24.

[26] A.V. Maltseva, O.V. Makhnytkina, N.E. Shilkina, "Studying of social media users' behavior patterns: using big data", in *The VI International Sociological Grushinsky Conference "Research life after the research: how to make results clear and useful"*, Moscow, 2016, pp. 988-991.

[27] M. Orešković, J. Benić, M. Essert, "A step toward machine recognition of complex sentences", *TEM Journal*, vol.7 (4), 2018, pp. 823-828.

[28] M. Oreškovic, M. Cubrilo, M. Essert, "The Development of a Network Thesaurus with Morpho-Semantic Word Markups", in *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, Tbilisi, 2016, pp. 273-279.

[29] N.V. Pushkareva, "Covert Sense in Prose Texts, Semantic and Linguistic Methods of its Implementation", *Vestnik of Saint Petersburg University. Language and Literature*, vol. 1(1), 2009, pp. 59-65.

[30] G.Y. Solganik "About a text modality as a semantic basis of the text", in: *The VIIIth International conference "Structure and semantics of the art text*, Moscow, 1999, pp. 364-372.

[31] N.E. Petrov *About the content and volume of a language modality,* Novosibirsk: Sciense, 1982 (in Russian)

[32] A.P. Beltyukov, M.M. Abbasi "Logical analysis of emotions in natural language texts", *Vestnik Udmurtskogo Universiteta: Matematika, Mekhanika, Komp'yuternye Nauki,* vol. 29(1), 2019, pp. 106-116.

[33] T.V. Romanova, *Modality. Assessment. Emotionality*, Nizhny Novgorod: NSLU N.A. Dobrolyubov, 2008.

[34] A. Maltseva, A. Klebanov, N. Shilkina, I. Lyamkin, O. Mahnitkina, "Culture of social media interactions amongst modern students: analysis of the social network vk.com, university groups «Overheard...» with big data", in *Proceedings of the International Conference IMS-2017*. ACM, New York, 2017, pp.11-14.

[35] V. V. Vinogradov *Lexicology and lexicography: selected works,* Moscow: Sciense, 1977.

[36] *Labor Code of the Russian Federation with comments and changes in 2018*. Available: http://tkodeksrf.ru.

[37] A.V. Maltseva, O.V. Makhnytkina, N.E. Shilkina, F.I. Mirzabalaeva, S.A. Ilyinykh, "Database of verbal markers about professional and labor intentions of student's youth (labexp)", Registration certificate RUS 2018620499, February 06, 2018.