# The sentiment-analysis algorithm of social networks text resources based on ontology

**N.G. Yarushkina**[1], **V.S. Moshkin**[1], **I.A. Andreev**[1]

[1]Ulyanovsk State Technical University, Severny Venetz street 32, Ulyanovsk, Russia, 432027

**Abstract.** In this paper the features of semantic and sentiment analysis of textual data of social networks are presented, and an original model and algorithm for sentiment analysis of textual fragments of social networks using fuzzy linguistic ontology are proposed. This approach involves the use of various subgraphs of fuzzy ontology when considering texts of various subject areas with regard to contexts. In addition, the algorithm involves the assessment of the sentiment scores of individual syntagmatic structures into which the analyzed text fragments are divided. It also presents the results of experiments comparing the efficiency of the developed algorithm with a group of existing approaches in analyzing text fragments on the example of data from the social network VKontakte.

## 1. Introduction

Social networks are an integral part of modern society. The audience of social networks is huge. The most popular social network in the world, Facebook, has more than a billion unique visitors per month. Twitter comes second with more than 300 million unique visitors. The most common VKontakte portal in Russia and the CIS countries has 80 million visitors per month. All users daily leave a huge number of messages reflecting the position of citizens from different countries and different sectors of society [1].

The study of social networks every year is becoming increasingly important because of the need to ensure the safety of the population and the monitoring of public sentiment. Analyzing posts and posts can help assess changes in users 'political and social attitudes.

Any commercial enterprises producing any product, it is important to know the opinion of buyers about this product. These data can be used to improve the quality of the product, determine the target audience and to identify the main advantages and disadvantages of competitors. This problem is solved by the Opinion mining. This analysis consists of two subtasks: 1) morphological analysis to identify entities that will be evaluated, and 2) analysis of the sentiment of expressions related to this entity.

By sentiment analyzing of the users' text messages the researcher can draw conclusions about:

- emotional evaluation of users of various events and objects;
- individual user preferences;
- some features of the users' nature [2].

The sentiment is the emotional attitude of the author to some object in the form of text. Currently, there is a set of methods for analyzing the tonality of textual information. First of all, these are methods based on the use of dictionaries and on machine learning with a teacher.

The main component of the concept of Web 2.0 is the development of social networks. Web 2.0 assumes the formation of the electronic resources content (including text) by users through their profiles (posts, comments, file names, file signatures of various formats, etc.). Therefore, text data in social networks have the following features:

- Use of slang turns, neologisms and also various dialectic forms.
- Use whole and incomplete sentences.
- The presence of speech and spelling errors.
- The use of smiles, emoji to give the message a certain emotional coloring.

In this article we will consider the use of various existing algorithms for assessing the sentiment of social network texts within the framework of the developed software system for Opinion Mining. The article proposes an original ontological method that takes into account the features of the text data presentation in social networks.

## 2. The groups of methods for sentiment analysis of text data

There are two main groups of methods for the automatic sentiment analysis of text data:

### *2.1. Statistical methods*

The basis of these methods is the use of machine classifier. This classifier is learned on pre-marked texts in the first stages. Then the classifier builds a model for analyzing new documents using the knowledge gained. The algorithm consists of:

- A collection of documents is collected for machine classifier learning;
- Each document is decomposed into a feature vector;
- The correct sentiment type is indicated for each document;
- The selection of the classification algorithm and the method for learning the classifier;
- The resulting model is used to determine the documents sentiment of the new collection.

The disadvantage of such methods is the need for a large amount of data for learning.

The statistical approach widely uses the support vector method (SVM) [3], Bayesian models [4], various types of regression [5], methods Word2Vec, Doc2Vec [6], CRF [7], convolutional and recurrent neural networks [8].

### *2.2. Methods based on dictionaries*

Tonal dictionaries and rules are compiled using linguistic analysis. These dictionaries search emotive words and expressions. Further, the set of emotive words is assessed on a scale containing the number of negative and positive vocabulary [9].

Methods based on dictionaries assume the presence of a linguist expert who compiles an exact reference book of emotionally colored words and expressions. Each expression (word, smile or style) is called a marker. Emotion is taken into account in the algorithm when finding the marker. The result of the algorithm is the average emotional color of the text [10-11]. The following algorithm is usually used:

- Assign the sentiment score from the dictionary to each word in the text;
- Calculate the overall sentiment score of the entire text by adding the sentiment score of individual words [12].

The disadvantages of this method is a significant amount of labor because the method requires the creation of many of rules.

In addition, sometimes used a mixed method (a combination of the first and second approaches) [13-14].

### *2.3. Ontological method*

The ontological method of sentiment analysis of textual data from social networks was developed as part of the study. This method is a modification of methods based on lexical dictionaries.

The general scheme of the developed ontological algorithm is presented in Figure 1.

The main features of the ontological method are:

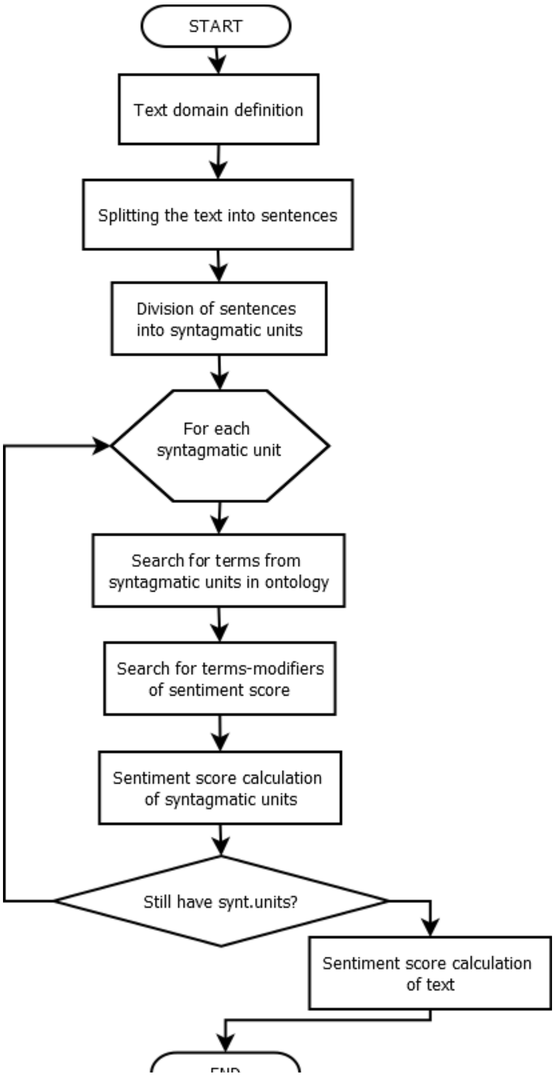
START
Text domain definition
Splitting the text into sentences
Division of sentences into syntagmatic units
For each syntagmatic unit
Search for terms from syntagmatic units in ontology
Search for terms-modifiers of sentiment score
Sentiment score calculation of syntagmatic units
Still have synt.units?
Sentiment score calculation of text
END

**Figure 1.** General scheme of the ontological method.

- Consideration of features of the subject area. The sentiment score of each syntagmatic unit always depends on the nature of the subject area. In Fig. 2 Pr1, Pr2 are subject areas, S1- S6 are the sentiment scores of the terms.

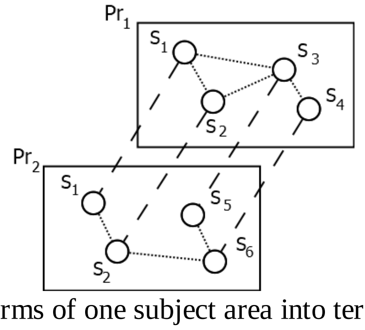

**Figure 2.** Translation of terms of one subject area into terms of another subject area.

- Considering the sentiment of syntagmatic units rather than individual words. A syntagmatic unit is an aggregate of several words united according to the principle of semantic grammatical phonetic compatibility [15].
- Classification of the subject area according to the set of terms used.

- Accounting for semantic relationships between objects (terms). First of all the relations of synonymy, hyponymy, etc. are taken into account. This functionality helps to assess the sentiment of slang turns, neologisms used in the texts of social networks and take into account spelling and speech errors.
- Consideration of the influence of terms that change the level of syntagmatic unit emotional coloring. Certain terms have a priori coefficients that change the overall sentiment score of a syntagmatic unit in combination with other terms. For example, words with amplification factors: “very”, “above”, “stronger”, etc., reductions - “less”, “lower”, negation - “not”, etc.

$$O = \{Synt^{PrA}, PrA, \rho_V^{Synt}, V^{Synt}\},$$

where $Synt^{PrA}$- a set of syntagmatic units of the analyzed text fragment; PrA - is the subject area, $\rho_V^{Synt}$ - is the set of sentiment score of the syntagmatic text units, where

$$\rho_V^{Synt} = k \times \rho_V,$$

where k - is the a priori coefficient of modification of the syntagmatic unit sentiment score, $\rho_V$-is the sentiment score of the term of the subject area, $\rho_V^{Synt} \in [0,1]$, $V^{Synt}$ – sentiment classes, $V^{Synt} = \{+, -\}$.

Hence, the overall sentiment score of the text

$$S_j = \sum_{j=1}^{n} Synt_j^{PrA_i}$$

$$S = \frac{S_j}{S_{max}} \times 100\%,$$

where $Synt_j^{PrA_i}$- is the j-th syntagmatic unit of the i-th subject area PrA [1..n], $S_K$ is the unnormalized value of the sentiment score for the k-th text fragment, $S_{max}$-is the maximum value of the of the sentiment score of a analyzed text fragment.

## 3. Experiments

420 posts and comments of groups and users of the social network VKontakte on the following topics were analyzed:

- Foreign policy;
- Musical groups;
- Film premieres;
- IT industry;
- Medicine and education;
- Activities of non-profit organizations.

These subject areas were chosen as there is always an assessment of real-world objects and the opinion of the author in these texts.

A lexical ontology was developed to evaluate the effectiveness of the ontological algorithm. The basis of this ontology is the dictionary of evaluative words SentiWordNet [16]. SentiWordNet is a lexical resource for analyzing opinions. SentiWordNet assigns three senses to each WordNet lexical unit: positivity, negativity, objectivity. The current official version of SentiWordNet is 3.0 based on WordNet 3.0 [17].

SentiWordNet has been complemented by Russian-language terminology. In addition, the definition of the subject area was added to the description of individual lexical units of the basic dictionary. Several records were created for the term, indicating different subject areas and the corresponding sentiment scores if the emotional component of the term in different subject areas had a different values. Thus, the terms in the created lexical ontology have the following order of description:

- ID;
- Term (English);
- Term (rus.);

- Subject area;
- The positive sentiment score [0;1];
- The negative sentiment score [0;1];
- Interpretation of the term (English);
- Interpretation of the term (rus.);
- Example of use (English);
- Example of use (rus.);
- Set ObjectProperties (synonymy, hyponymy, etc.).

Learning set for machine learning algorithms was text data from the social network VKontakte in Russian. All sentences in this sample end with a smile expressing joy - :) or sadness - :(. Smiles are markers that define the a priori sentiment of sentences without peer review.

A number of experiments were conducted to evaluate sentiment scores of the text taken from social networks using classification algorithms based on machine learning:

- Naive Bayes classifier.
- Linear regression.
- Support vector machine.
- LSTM network.
- The developed method based on ontology, obtained from SentiWordNet.

The SciKit Learn library was used to evaluate the effectiveness of the first three methods. The TensorFlow library was used to simulate the operation of the LSTM network. All libraries are written in Python. The sentiment of each sentence was previously estimated by three experts. The conclusion was made about the accuracy of the sentiment analysis by the corresponding algorithms on the basis of expert estimates. The results of the experiments are shown in table 1.

**Table 1.** Results of experiments.

| № | Algorithm | Learning parameters | The number of correctly defined sentences | % |
|---|---|---|---|---|
| 1 | Naive Bayes classifier | Unigrams | 329 | 78,33 |
| | | Bigrams | 324 | 77,14 |
| 2 | SVM | Unigrams | 308 | 73,33 |
| | | Bigrams | 316 | 75,25 |
| 3 | Linear regression | Unigrams | 274 | 65,24 |
| | | Bigrams | 270 | 64,3 |
| 4 | LSTM | | 320 | 76,19 |
| 5 | Ontology algorithm | | 323 | 76,9 |

The best result was shown by the method of machine learning with the use of the Naive Bayes classifier. The Bayesian classifier was trained on unigrams and bigrams: in both cases, the classification results were better than all other algorithms.

The developed method based on the dictionary obtained from SentiWordNet showed an efficiency of 77%. Recurrent neural networks showed an efficiency of 76%.

The result of the work of algorithms based on machine learning is highly dependent on the training set. The idea of evaluating the texts sentiment score by smiles at the end of sentences justified the expectations.

## 4. Conclusion

Thus, in this paper, a set of methods of sentiment analysis of Russian-language texts of social networks in Russian was analyzed as part of a project to develop an automated system for analyzing opinions. A number of experiments were conducted to evaluate the sentiment score of text messages on the social network VKontakte using a set of statistical methods and the developed method using fuzzy lexical ontology obtained from the SentiWordNet 3.0 dictionary.

The developed method using fuzzy lexical ontology showed a good result - efficiency in the region of 77%. Our research team plans to modify this method in the following areas:

- extension of the terminology of lexical ontology, due to neologisms and dialecticism;
- extension of the set of semantic relations between terms in ontology;
- modification of the model for calculating the overall sentiment score of the sentence by taking into account the significance of each term included in it;
- development of algorithms for automatic extension of the core of the developed lexical ontology using morphological and semantic analysis of social networks texts.

## 5. Acknowledgement

This study was supported by the Russian Foundation for Basic Research (Grants No. 18-47-732007, 18-47-730019 and 18-47-730035).

## 6. References

[1] Shipilov, O.Yu. Analysis of the emotional color of messages in the social network twitter / O.Yu. Shipilov, A.S. Belyaev // Science Questions. – 2016. – Vol. 3. – P. 91-98.

[2] Vlasov, D.A. Description of the information image of a social network user, taking into account its psychological characteristics // International Journal of Open Information Technologies. – 2018. – Vol. 6.

[3] Sabuj, M.S. Opinion Mining Using Vector Machine for Web Based Diverse Data / M.S. Sabuj, Z. Afrin, K.M.A. Hasan // Pattern Recognition and Machine Intelligence. Lecture Notes in Computer Science. – 2017. – Vol. 10597. – P. 673-678.

[4] Dinu, L.P. The Best Feature of the Set / L.P. Dinu, I. Iuga // Computational Linguistics and Intelligent Text Processing. CICLing. Lecture Notes in Computer Science. – 2012. – Vol. 7181. – P. 556-567.

[5] Chetviorkin, I.I. Sentiment Analysis Track at ROMIP-2012. Computational linguistics and intellectual technologies. Computational linguistics and intellectual technologies: "Dialogue-2013" / I.I. Chetviorkin, N.V. Loukachevitch // Sat scientific articles. – 2013. – Vol. 2. – P. 40-50.

[6] Chen, Q. Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries / Q. Chen, M. Sokolova // CoRR abs. – 2018. – 1805.00352.

[7] Antonova, A. Using the conditional random fields method for processing texts in Russian / A. Antonova, A. Soloviev // Computational linguistics and intellectual technologies: "Dialogue-2013". Sat scientific articles. – 2013. – Vol. 12(19). – P. 27-44.

[8] Maas, A.L. Learning word vectors for sentiment analysis / A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts // The International Language Technologies – International Association for Computational Linguistics. – 2011. – Vol. 1. – P. 142-150.

[9] Saif, H. Contextual semantics for sentiment analysis of Twitter // Information Processing & Management. – 2016. – T. 52, № 1. – P. 5-19.

[10] Pak, A. Twitter as a Corpus for Sentiment Analysis and Opinion Mining / A. Pak, P. Paroubek // LREC, 2010.

[11] Tarasova, A.N. Synergy of interrogative and exclamation marks in network texts (on the material of Tatar, Russian and English languages) // Bulletin of Vyatka State University, 2015.

[12] Ionova, S.V. Emotiveness of a Text as a Linguistic Problem // Abstract. Diss. Cand. filol. Sciences, 1998.

[13] Pang, B. Thumbs up? / B. Pang, L. Lee, Sh. Vaithyanathan // Sentiment Classification using Machine Learning Techniques. – 2002. – P. 79-86.

[14] Turney, P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proceedings of the Association for Computational Linguistics. – 2002. – P. 417-424.

[15] Zarubin, A. Session Data Science / A. Zarubin, V. Moshkin, A. Filippov, A. Kovalov – Samara, Russia, 2018. – P. 179-185.

[16] Esuli, A. SENTIWORDNET: A Guide for Respecting the Opinion Mining / A. Esuli, F. Sebastiani, 2006. – P. 417-422.

[17] Miller, G.A. WordNet: A Lexical Database for English / G.A. Miller // Communications of the ACM. – 1995. – Vol. 38(11). – P. 39-41.