Topic Modeling Application for Intellectual Analysis of Reviews in Russian*

Katsiaryna V. Kosarava[1][0000-0001-7326-5307] and Natalia V. Davydik[1][0000-0002-5735-135X]

[1] Grodno State University of Yanka Kupala, Orzeshko str. 22, 230022 Grodno, Belarus
kosareva.ev@mf.grsu.by

Abstract. In the article, the problem of applying the methods of intellectual analysis for processing the texts of reviews of websites in Russian is investigated. The stages of data collection and preprocessing are described. LDA model for determining the topics of the collected reviews is described and trained. Some possible, approaches for improving the constructed model are described: the usage of the TF-IDF algorithm, increasing the size of the dataset, and usage of the Mallet algorithm. The performance of the models improved using the described methods was evaluated based on the coherence coefficients. The visualization graphs of the constructed topics are built using Python library for interactive topic model visualization pyLDAvis. The optimal number of topics and the distribution of keywords for the constructed topics were determined. An example of the best LDA model topic prediction for new reviews is given.

Keywords: Topic Modeling, Natural Language Processing, Dirichlet Latent Placement Method.

1 Introduction

For the efficient operation of any company, one of the important components is competitiveness and a stable image in the eyes of its customers. And when choosing a certain service or product of a company, a consumer, having no experience with this organization, will pay priority attention to the experience of other clients from this organization. The client can gain this experience by reading the reviews on the company's website. And based on these reviews, it forms its attitude towards the organization that provides goods or services. Automation of processing reviews is especially important for large on-line trading platforms, such as Amazon, Ebay.com, Wildberries, Ozon, because this approach allows you to identify the weaknesses of the company and makes it possible to improve them. After all, the effort to meet the needs of customers can increase the company's competitiveness. One of the ways to automate the processing of reviews is the usage of intelligent text analysis methods, such as topic modeling.

This approach will save time processing, continuously storage data permits to determine the review subjects more correctly and find more rational solutions to the problems.

2 Topic Modeling

2.1 Overview of Available Technologies

As an example, public services' on-line text analysis can consider the "Tone Analyzer" [1] and "Analysis letters" [2]. The Tone Analyzer service is English-language and uses linguistic analysis to detect emotional and language tones in written text. This system can analyze the tone of both a document and a single sentence [1].

The Tone Analyzer service is used to evaluate how written messages will be perceived, and then it is possible to improve their tone. Businesses can use this service to learn the tone of the message from their customers and respond appropriately to every opinion.

The second service considered is "Mail Analysis" - a Russian-language online service that analyzes the mood and emotional state of a person based on the text of his email message. This service determines not only the mood of the letter author but also interprets the analysis result [2].

However, the considered services provide an opportunity to determine whether the analyzed text is positive or negative but do not reveal the topic and the problem. And understanding what exactly did not suit the client makes it possible to improve the activities of the enterprise.

2.2 Automated Processing of Reviews Using Artificial Intelligence Techniques

Artificial intelligence techniques are becoming widespread due to the need to process large amounts of data. And some of the most promising and common methods are machine learning methods, the main advantage of which is the constant learning, development, and expansion of the vocabulary of the neural network. Natural Language Processing (NLP) is a general area of artificial intelligence and mathematical linguistics. It studies the problems of computer analysis and synthesis of natural languages. When applied to artificial intelligence, analysis means understanding a language, and synthesis means generating a literate text. Solving these problems will mean creating a more convenient form of interaction between a computer and a person. NLP can be widely used in such tasks as creating recommendation systems based on a textual description of the content, searching for a suitable educational on-line course or vacancy on the Internet, etc[3].

The introduction of an automated system for processing feedback using the artificial intelligence techniques will reduce the processing time of feedbacks by an employee of the company (identifying the problem), promptly respond to negative feedbacks and resolve conflict situations, save time when automatically sending a letter to the appropriate specialist and receive detailed analytics about the company's image.

The main task of automating the review process is to identify and extract keywords (topics). Based on these words, an idea about the topic of the client’s reviews is formed.

This enables the company's employees to respond to these reviews on time and propose a solution to the problem.

The most common way to define a topic of the text is topic modeling, which is based on building a topic model. This model receives as an input a collection of documents presented in text form. Model output is a vector of numbers that characterize the degree of belonging of the document to certain topics. The dimension of the vector is formed either through the number of topics that were specified at the input of the model and, or generated automatically by the model itself [4].

One of the most commonly used methods for determining topics is the Latent Dirichlet Placement Method (LDA). This algorithm is used to detect topics that are present in the document. The LDA method is an improved version of the Probabilistic Latent Semantic Method (PLSA) and includes some additions regarding the document vector. Also, according to the LDA model, in contrast to the PLSA model, it is believed that the topics are distributed directly according to the Dirichlet distribution and in practice, a more correct set of topics is formed. The LDA algorithm is applied after cleaning and vectoring the text. When the text passes through this algorithm, the distribution of words for each topic and the distribution of topics in each document is determined.

Using LDA, you can extract human-interpreted topics from a corpus of documents, where each topic is characterized by the words which are most strongly associated with this topic.

There are many open-source libraries for topic modeling, but one of the most widely used libraries for the Python programming language is Gensim. The Gensim library allows you to process texts, generate thematic models, and scales very well for large text corpora, however, it does not include factorization of non-negative matrices, which can also be used to search for topics in the text [5].

2.3 Dataset Description

To build and train the LDA model, a dataset was collected from the website otzovik.ru. It is consisting of 2500 reviews in Russian on automotive topics, Table 1. Each review includes text and rating. To gather feedbacks «Web Scraper» for browsers was used. It allows you to obtain data by extracting it from various Internet resources. This web scraper does not require deep programming skills from the user; the scraper is configured by selecting the necessary elements on the page. This method can easily extract data from dynamic websites with multiple levels of navigation. It can also navigate the website at all levels. After collecting data, the web scraper generates a dataset in CSV (delimited data) format, which can be downloaded directly from the browser.

Table 1. Table captions should be placed above the tables.

Rating	5	4	3	2	1
Reviews number	1007	413	253	243	591

After the dataset with reviews has been collected, it is necessary to preprocess it. Text preprocessing includes the following steps: converting text to lower case, removing punctuation, numbers, and special characters, tokenization, stop words removing, stemming, and lemmatization. To perform the first four stages of text preprocessing the

NLTK (Natural Language Toolkit) library was used. NLTK is the leading platform, which contains a set of modules for creating natural language processing programs in Python. This library has easy-to-use interfaces for more than 50 corpora and lexical resources [6]. Besides NLTK has a Russian stopword list and a punctuation list that you can import.

One of the main stages of text preprocessing is converting words to their initial form. In this work, the lemmatization method was used. This method takes into account the meaning of the text and carries out morphological analysis when reducing the word to its canonical form [7]. Since the dataset of reviews was collected in Russian, the morphological analyzer pymorphy2 was used. This library works with dictionaries from the open corpus of the Russian language OpenCorpora and builds hypotheses for unknown words. At the same time, it does not use data on neighboring words when the morphological analysis is carrying out [8]. For a more detailed analysis of a word, you can use tags consisting of a set of grammatical meanings describing the given word. The MorphAnalyzer.parse() method returns a large number of all possible word parses.

2.4 LDA Model Training

To carry out topic modeling, it is necessary to extract topics from the text of the reviews that will characterize the customer calls. In the future, the solution to the client's problem situation will be carried out by an employee of the company, depending on a specific topic. The most popular algorithm for topic modeling is Latent Dirichlet Placement (LDA). This algorithm is implemented in the Gensim library, which is best suited for unsupervised model training, topic modeling, and natural language processing. For this experiment, the highlighted features of the Gensim library are decisive, since the model is trained independently without intervention from the teacher [5].

In this case, the very concept of "topic" is a collection of keywords describing this topic. To get a good result in highlighting key topics, the initial training data should contain a wide variety of topics. The choice of the topic modeling algorithm and the correct settings of this algorithm are also important.

Before building the LDA model, it is necessary to form the main input data for the implementation of topic modeling: a dictionary and a corpus. The dictionary encapsulates the mapping between normalized words and their integer identifiers (id2word). The Gensim package provides the ability to create a unique identifier for each word. Corpus is a stream of vectors of a document or a sparse matrix, which looks like this (id_word, word_frequency). So, the corpus describes the frequency of occurrence of each word in the text. In this case, the doc2bow format is used - this is the “bag of words” model, which is one of the ways to vectorize and simplify the representation of words for their further processing [5]. If the corpus is not specified, then the model remains untrained, since it is assumed that the model will be trained with the help of a teacher. In addition to the corpus and vocabulary, it is necessary to specify such parameters as the number of topics that are distributed in the training corpus, the number of documents that will be repeated for each update, and the number of documents that are used in each training block.

The LDA model described in this article is based on 2,500 reviews and highlights from 2 to 10 topics. Each topic is a combination of keywords. The words with the greatest weight are defining for this topic. Looking at the keywords can help you get an idea of the topic.

When constructing an LDA model, it is also necessary to consider topic consistency (coherence). This indicator measures how well the topic modeling has been performed. Coherence measures how often the most likely words of a topic appear side by side in a training dataset. To find the optimal number of topics, it is necessary to build many LDA models, which will contain a different number of topics. The model with the highest consistency value will be optimal. If the same keyword is repeated in several topics, this indicates that the selected number of topics is too large. However, the coefficient of coherence only gives an understanding of the frequency of occurrence of keywords together for the selected topics. To understand how best to group topics for the further practical application we should use topic visualization.

For a data set consisting of 2,500 reviews in Russian, the largest value of the coefficient of coherence 0.3333 was obtained for the LDA model with 10 topics. Fig. 1 shows the values of the obtained consistency coefficient for a different number of topics.

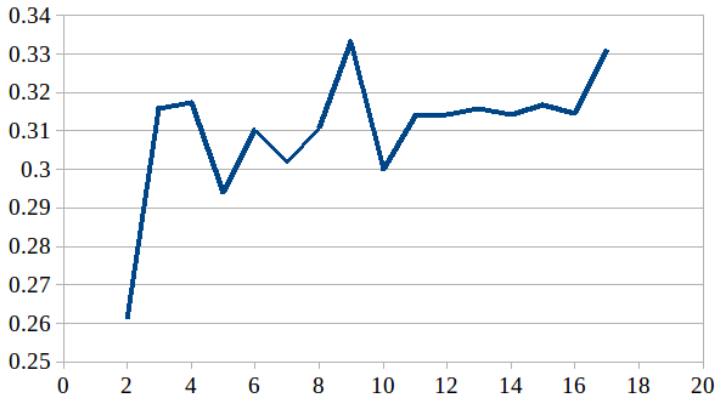

Fig. 1. Coherence per topic number.

An interactive diagram of the package pyLDAvis was used to visualize topics and keywords in the LDA model with 10 topics [9]. Circles ("bubbles") - a global view of the constructed model. The centers of these circles are determined by calculating the distance between topics. The area and scale of the bubbles reflect the prevalence of each topic. That is the larger the circle, the more common that topic is in the document. And all the bubbles are arranged in descending order of the prevalence rates. A well-constructed topic model should have large, non-intersecting circles that are represented across the entire plane of the diagram, rather than grouped in one place. If the model has small "bubbles" located in one area of the diagram, which has a lot of overlap, this indicates that a lot of topics were highlighted [9].

According to the results of the experiment, the best visualization and distribution of topics was achieved when selecting 4 topics, even though the coherence coefficient in the model for 10 topics was higher. Fig. 2 shows a visualization of the LDA model for 4 topics. It is noticeable on the graph that circles 1 and 2 are large, located far from each other, and have no intersections, and "bubbles" 3 and 4 are smaller than the previous ones and are located close, but do not intersect. This suggests that this model is suitable for practical use and grouping customer calls by topic. However, there is an assumption that with further improvement of the model, themes 3 and 4 can be combined into one.

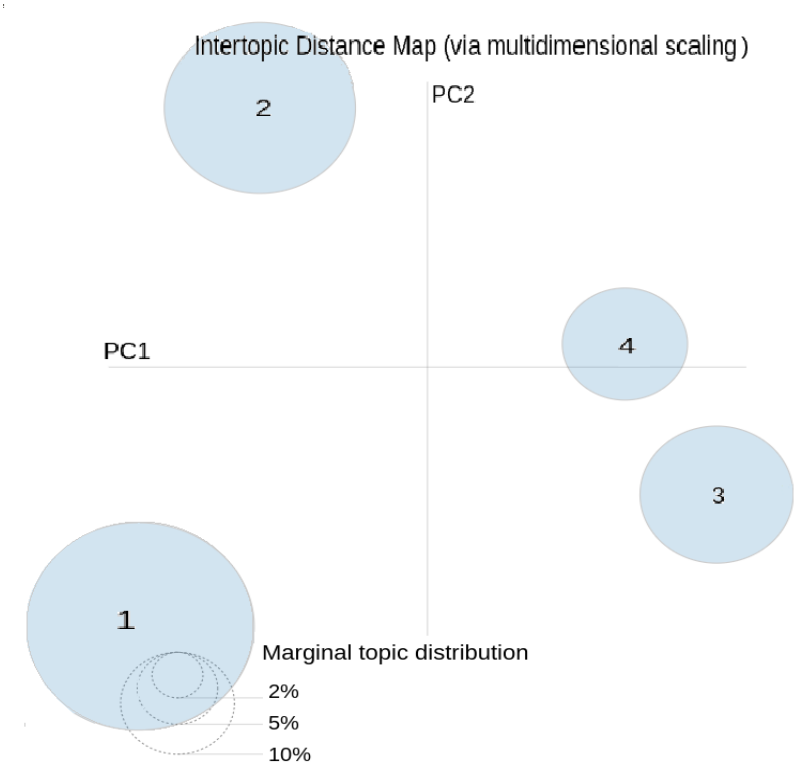

Fig. 2. Visualization of the LDA model for 4 topics.

2.5 Some Approaches to Improve Model Performance

Since some words occurring in the dataset may not be included in the list of stop words and may not be informative for training and building the LDA model, it is possible to expand the list of stop words. To do this, you need to evaluate the significance of the words of the constructed dictionary for the entire corpus of reviews and each review separately. The statistical measure Tf-Idf can be used as an indicator of significance. Tf-Idf is using to estimate the importance of the word in the context of the document,

which is part of the document corpus. The weight of a word is proportional to the frequency of using this word in a document and inversely proportional to the frequency of using a word in all documents of the corpus [10]. Thus, the more often a word occurs in a separate document and rarely in the entire corpus, the greater Tf-Idf value will be. Conversely, the more often a word occurs in all documents in the corpus, the lower its Tf-Idf value will be. Thus, words with a small Tf-Idf value are specific to a given set of reviews and can be added to the stop word list.

Another way to improve the model performance is to increase the size of the training dataset. To test this assumption, about 2,000 more reviews on automotive topics, (which include information about auto parts, car acoustics, car care products, etc.) were collected As a result, a dataset of 4,500 reviews was generated. As well as for the model built on 2500 reviews, the greatest topic consistency was obtained for the model with 10 topics. The value of the coherence coefficient is 0.3637 and it is larger than the one obtained in the first model trained on 2500 reviews. Thus, it can be concluded that increasing the size of the dataset increases the topic consistency. The visualization showed that the best LDA model was built for 3 topics and the coherence coefficient was 0.3295. For comparison, a coherence coefficient of the Another factor affecting the performance of the model is the correct choice of the training algorithm. The gensim library implements the LDA Mallet algorithm [11], which often allows better highlighting of topics and its use increases the quality of the model. Since the LDA model trained on a dataset of 4,500 reviews showed the best result, further improvement using the LDA Mallet algorithm was carried out on this model. According to the results of the experiment, the optimal number of topics was 3. Although the coherence coefficient decreased to 0.2776, the pyLDAvis visualization showed the best separation of topics and their prevalence. Fig.3 displays the LDA Mallet model with 3 topics trained on a dataset of 4500 reviews.

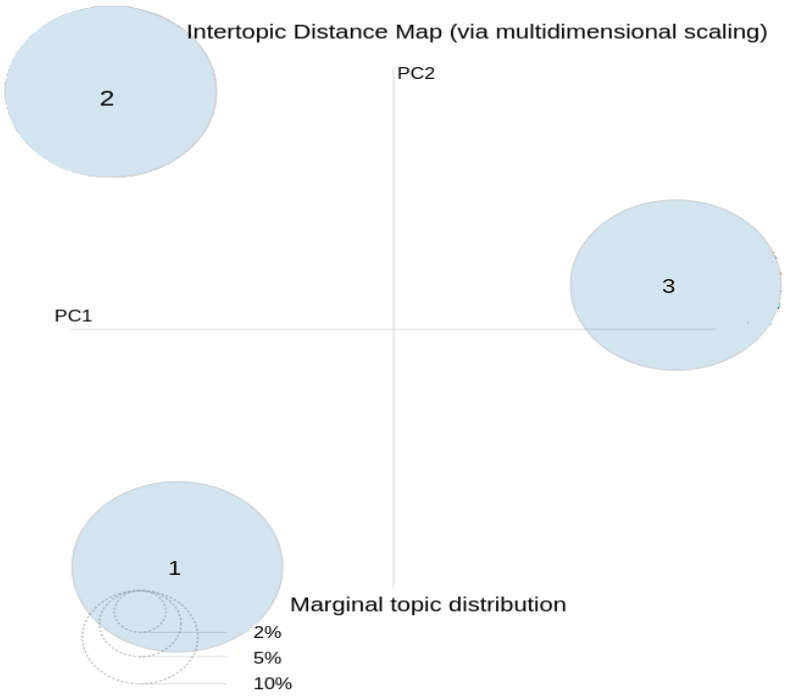

Fig. 3. Visualization of the LDA Mallet model with 3 topics trained on 4500 reviews.

The graph shows that all the circles have the same size, are located at a great distance from each other, and don't intersect.

3 Experiment Results

Let's consider the results obtained for the LDA Mallet model trained on a dataset of 4500 reviews, which was described in the previous section. Figure 4 shows the top 10 words (translated from Russian) describing the first topic in the described model.

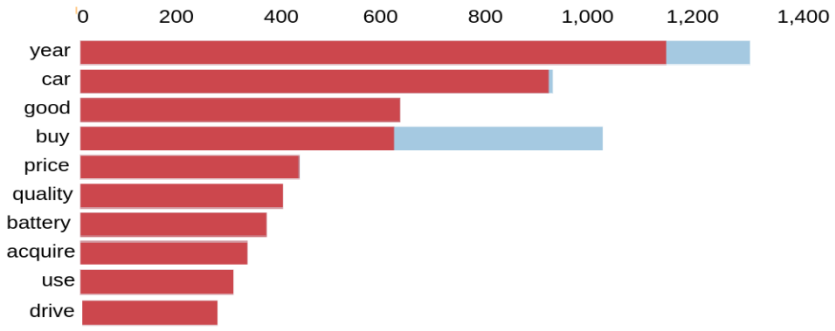

Fig. 4. Top 10 keywords for the first topic.

On the left of the graph, keywords are marked, and on top - their number in the first topic. Based on the displayed keywords, we can conclude that the first topic is related to buying a car since keywords characterize the main questions that people ask when purchasing a car, namely: year, price, quality, as used by the previous owner (at the same time, the words buy and acquire are clearly defined).

Let's consider the top 10 words for the second and the third topics. The most popular words in the second topic are "decide", "auto" and "work", which appear in reviews more than 300 times. The last words from the top 10 are "store", "block", "company", "put", "manufacturer", "return", "become", they appear in reviews more than 200 times. Based on the keywords we can conclude that this topic is related to the client's consultation since the most common keyword for this topic is "decide". Probably, this topic expresses the client's intention to get advice on the choice of a car, auto parts from the manufacturer, or the work of the store.

The top 10 words from the third topic are "vehicle", "oil", "auto", "tire", "bearing", "work", "given", "system", "big", "problem". Based on these keywords, it can be assumed that the third topic is related to auto parts and customer questions are addressed to the auto service of the car manufacturer since the main keywords describe the operation of the care system, parts, and components. Figure 5 shows the distribution of weighted keywords in the resulting topics.

```
      [(0,  '0.041*"year"  +  0.032*"car"  +  0.029*"good"  +
0.013*"buy" + 0.011*"price" + 0.011*"quality" +  0.010*"battery"
+ 0.008*"acquire" + 0.008*"use" + 0.008*"drive"')

      [(1,  '0.041*"decide"  +  0.021*"auto"  +  0.020*"work"  +
0.015*"store" + 0.014*"block" + 0.012*"company" +  0.011*"put" +
0.011*"manufacturer" + 0.010*"return"+ 0.010*"become"')]

      [(2,  '0.015*"vehicle"  +  0.013*"oil"  +  0.013*"auto"  +
0.012*"tire" + 0.011*"bearing" + 0.010*"work" + 0.009* "given" +
0.009*"system" + 0.009*"big" + 0.009*"problem"')]
```

Fig. 5. Keywords distribution in topics.

Let's test the trained model on new data. Table 2 shows the results of the model using the example of six reviews. The first column of the table contains the text of the review, the second - the vector generated by the LDA Mallet model, which determines the probabilities of topics (rounded to two decimal places). The third column displays the name of the topic the review was assigned to.

Table 2. Results of topic modeling for new reviews.

Appeal/feedback text	Vector of belonging of reviews to the topic	Topic
Good day. I am interested in the engine power in the Atlas model	(0.34, 0.36, 0.30)	Consultation
Hello. I want to sign up for a test drive. I live in Grodno.	(0.35, 0.33, 0.33)	Buying a car/test drive
I want to buy a car, could you advise me and send me a price list.	(0.30, 0.38, 0.32)	Consultation
Good evening. I am interested in the components for the EMGRAND model: pads and tires, winter tires	(0.30, 0.34, 0.35)	Auto parts
Tell me which manufacturer to choose for antifreeze.	(0.33, 0.36, 0.31)	Consultation
Yesterday, new oil was poured into the service, the quality leaves much to be desired, I'm not satisfied at all.	(0.31, 0.33, 0.36)	Auto parts

As we can see from table 2, the LDA Mallet model correctly identified the topics for new customer reviews, and based on the selected keywords, the model indicated the weight of reviews belonging to each topic.

4 Conclusion

The article describes the implementation of the collection and preprocessing of review texts using artificial inelegance methods in the Python programming language. LDA models have been built and trained for a different number of reviews, some approaches for improving these models have been analyzed. The optimal number of topics and the distribution of keywords for the selected topics have been determined.

5 References

1. Tone Analyzer Homepage, https://tone-analyzer-demo.ng.bluemix.net
2. Analysis letters Homepage, http://www.analizpisem.ru/index.html
3. Kosarava, K., Bujnitskaya, E.: Application of Methods of Intellectual Analysis of Weakly Structured Data for Searching Online Courses. Proceedings of III International scientific and

practical conference “Scientific Interdisciplinary Research”, Saratov: SSO "Digital Science", 26-30 (2020).
4. Jones, S.: Automatic Extraction of Document Key Phrases for Use in Digital Libraries: Evaluation and Applications. Journal of the American Society for Information Science and Technology 8(53), 653-677 (2002), DOI: 10.1002/asi.10068.
5. Saxton, Micah D.: A Gentle Introduction to Topic Modeling Using Python. Theological Librarianship 1(11), 18-27 (2018), DOI: 10.31046/tl.v11i1.506.
6. Natural language processing for Python, https://www.nltk.org/book
7. Uysal, A., Serkan, Günal: The Impact of Preprocessing on Text Classification. *Inf. Process. Manag.* 1 (50), 104-112 (2014), DOI:10.1016/j.ipm.2013.08.006.
8. Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks, and Texts, 320-332 (2015), DOI: 10.1007/978-3-319-26123-2_31.
9. PyLDAvis Homepage, https://pyldavis.readthedocs.io/en/latest/modules/API.html.
10. Liu, Q., Wang, J., Zhang, D., Yang, Y., Naiyao, W.: Text Features Extraction Based on TF-IDF Associating Semantic. 2338-2343 (2018), DOI: 10.1109/CompComm.2018.8780663.
11. Ebeid, I., Arango, J.: Mallet vs GenSim: Topic Modeling Evaluation Report, (2006), DOI: 10.13140/RG.2.2.19179.39205/1