ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР
ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION AND COMMUNICATION TECHNOLOGIES

DOI 10.51885/1561-4212_2023_4_211
IRSTI 20.19.27

G. Zhomartkyzy[1], I. Manapov[1], M. Bazarova[2], A. Urkumbaeva[1], M. Rakysheva[1], Umoh Oto-obong Ezekiel[1]
[1]D. Serikbayev East Kazakhstan technical university, Ust-Kamenogorsk, Kazakhstan
E-mail: gzhomartkyzy@edu.ektu.kz
E-mail: manapov.ildar123@gmail.com
E-mail: aurkumbaeva@edu.ektu.kz
E-mail: mrakysheva@edu.ektu.kz
E-mail: umoh.o@edu.ektu.kz
[2]Sarsen Amanzholov East Kazakhstan University
E-mail: madina9959843@gmail.com

ASPECT ORIENTED SENTIMENT ANALYSIS OF USER TEXT MESSAGES

ПАЙДАЛАНУШЫЛАРДЫҢ МӘТІНДІК ХАБАРЛАМАЛАРЫН АСПЕКТІГЕ БАҒДАРЛАНҒАН ТАЛДАУЫ

АСПЕКТНО-ОРИЕНТИРОВАННЫЙ АНАЛИЗ ТЕКСТОВЫХ СООБЩЕНИЙ ПОЛЬЗОВАТЕЛЕЙ

Annotation. *Aspect-oriented sentiment analysis plays a crucial role in understanding users' opinions and sentiments towards specific product or service features. This study investigates intelligent algorithms for aspect-oriented tone analysis of user text messages, focusing on smartphone reviews as a case study. The study includes data collection and preprocessing, studying the methods and performance of models for aspect extraction and tone analysis. The performance of these models was compared on test datasets containing customer reviews. For aspect extraction, we combine cross-lingual syntactic analysis with topic models to improve accuracy over the combination of Russian-language syntactic analysis and topic models. In particular, the BERT transformer-based model, BER Topic, shows high performance in aspect detection due to its ability to understand the context in sentences. In the sentiment analysis task, the RuBERT-tiny model based on the BERT transformer outperforms the others, showing higher accuracy in sentiment classification in smartphone reviews. This study provides valuable insights into aspect-oriented tonality analysis, emphasizing the importance of selecting appropriate methods and approaches. The results provide researchers and practitioners with recommendations for effective aspect-oriented tonality analysis.*

Keywords: *Sentiment Analysis, Natural Language Processing (NLP), Machine Learning, Aspect Extraction, Text Analytics.*

Аңдатпа. *Аспектіге бағдарланған көңіл-күй талдауы пайдаланушылардың нақты өнім немесе қызмет сипаттамаларына қатысты пікірлері мен көңіл-күйлерін түсінуде маңызды рөл атқарады. Бұл зерттеу жұмысында пайдаланушы мәтіндік хабарларының аспектілерге негіз-делген тональдығын талдау бойынша интеллектуалды алгоритмдері зерттелді, мысал ретінде смартфондарға қалдырылған пікірлер қолданылды. Зерттеу жұмысы деректерді жинау және алдын ала өңдеуден, аспектілерді шығару және тональдылықты талдау үшін әдістер мен модельдер тиімділігін зерттеуден тұрады. Бұл модельдердің тиімділігі пайдаланушылардың пікірлерін қамтитын тестілік деректер жиынтығында салыстырылды. Аспектілерді шығару үшін тіларалық синтаксистік талдауды және тақырыптық модельдер біріктірілді, бұл орыс*

тіліндегі синтаксистік талдау мен тақырып модельдермен салыстырғанда дәлдікті жоғарлату мақсатында орындалды. Атап айтқанда, BERT трансформеріне негізделген BERTopic моделі сөйлемдердегі контексті түсіну қасиетіне байланысты аспектілерді анықтауда жоғары тиімділікті көрсетті. Көңіл-күйді талдау тапсырмасында BERT трансформері негізіндегі RuBERT-tiny моделі пікірлер тональдылығын классификациялауда жоғары дәлдікті көрсетіп, басқа модельдерден басым нәтижелер ұсынды. Бұл зерттеу сәйкес әдістер мен тәсілдерді таңдаудың маңыздылығын көрсете отырып, аспектілерге негізделген тональдылықты талдау туралы құнды деректерді ұсынады. Зерттеу нәтижелері зерттеушілер мен тәжірибе жүргізушілерге аспектіге негізделген тональдылықты тиімді талдау бойынша ұсыныстар ұсынады.

Түйін сөздер: *тоналдықты талдау, табиғи тілді өңдеу (NLP), машинамен оқыту, аспектілерді шығару, мәтінді талдау.*

Аннотация. *Аспектно-ориентированный анализ настроений играет решающую роль в понимании мнений и настроений пользователей по отношению к конкретным характеристикам продуктов или услуг. В этом исследовании производится исследование интеллектуальных алгоритмов для аспектно-ориентированного анализа тональности пользовательских текстовых сообщений, ориентируясь в качестве примера на отзывы на смартфоны. Исследование включает в себя сбор данных и их предварительную обработку, изучение методов и производительности моделей для извлечения аспектов и анализа тональности. Эффективность этих моделей сравнивалась на тестовых наборах данных, содержащих отзывы клиентов. Для извлечения аспектов мы комбинируем межъязыковой синтаксический анализ с тематическими моделями, чтобы улучшить точность по сравнению с комбинацией русскоязычного синтаксического анализа и тематических моделей. В частности, модель на основе трансформера BERT - BERTopic, демонстрирует высокую производительность при определении аспектов благодаря ее способности понимать контекст в предложениях. В задаче анализа настроений модель RuBERT-tiny на основе трансформера BERT превосходит других, демонстрируя более высокую точность классификации настроений в отзывах на смартфоны. Данное исследование дает ценную информацию об аспектно-ориентированном анализе тональности, подчеркивая важность выбора подходящих методов и подходов. Результаты дают исследователям и практикам рекомендации по эффективному аспектно-ориентированному анализу тональности.*

Ключевые слова: *анализ тональности, обработка естественного языка (NLP), машинное обучение, извлечение аспектов, анализ текста.*

Introduction. Today, product reviews can significantly influence the decision-making process of customers. They serve as an invaluable resource, offering information about product features, strengths and weaknesses, helping customers make informed purchasing choices. Positive reviews build confidence in a product, while negative reviews can potentially weaken it.

Analyzing the tone of customer reviews is an important task to understand how customers perceive and interact with products. The main purpose of tone analysis is to categorize the sentiment expressed in the text as positive or negative [1].

However, simply analyzing tone alone may not give a complete picture. A review may be negative in general, but positive about certain aspects of the product, because in most cases, it is not the object as a whole that people are expressing their opinions about. For this reason, the use of aspect-oriented tone analysis is useful.

Aspect-oriented tone analysis focuses on determining the sentiment expressed in relation to certain aspects of the object the text is about [2, 3]. An aspect can be a product characteristic, a service attribute, or any other aspect of the analyzed object. In the context of smartphone reviews, aspects can be various features or characteristics of a smartphone that the reviewer evaluates, such as camera, battery capacity, and memory.

Aspect-level tone analysis can help companies understand what customers like or dislike about their products and increase customer satisfaction, ultimately increasing sales. In addition, aspect-level tone analysis can help a company tailor product advertising to emphasize the strengths of the product.

Materials and methods of research The aspect-oriented tonality analysis consists of the following steps:

1) Raw Data Collection involves collecting unstructured raw data from various sources in their raw, unprocessed form for use in the ongoing aspect-oriented tone analysis. Smartphone reviews are used as the collected data in this study [4].

2) Pre-processing of collected data involves a series of steps applied to the collected raw data to convert it into a "clean", structured format suitable for further analysis [5].

3) Identifying aspects in the documents of the pre-processed dataset - involves identifying an aspect in each document of the pre-processed dataset. For this purpose, topic modeling techniques such as Latent Dirichlet Distribution (LDA), Gibbs sampled Dirichlet mixture model (GSDMM) or BERTopic are used. The output of this step is the aspect information contained in the documents of the pre-processed dataset.

4) Aspect-oriented analysis of the tone of certain aspects - focuses on assessing the sentiment expressed in relation to each aspect defined in the text.

Consider each stage of this study:

Collection of "raw" unprocessed data. For this study, smartphone reviews were collected from two different sources to ensure diversity in the data set. The first source was a well-known in Kazakhstan online store "White Wind", specializing in the sale of various electronics, including smartphones. Web page scraping was used to collect feedback from the pages of the online store [6]. In order to increase the diversity of the collected data, it is supplemented with reviews from the THEO VALL dataset from the Kaggle website. This dataset contains raw "raw" reviews from various online stores. The composition of the dataset is shown in Figure 1.

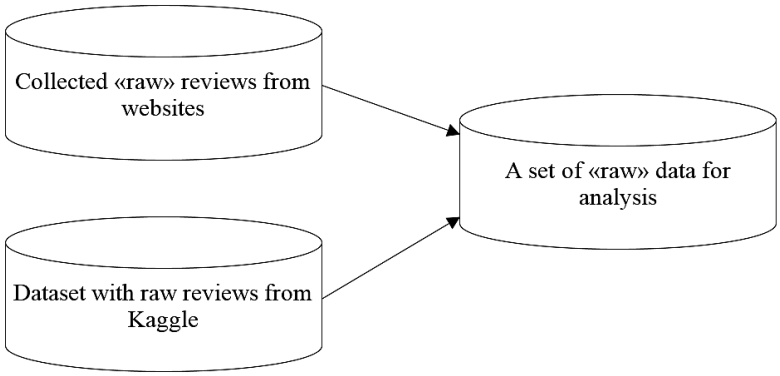

Figure 1. Composition of the dataset for aspect-oriented tonality analysis

1. Pre-processing of collected "raw" data [7, 8].

Preparing the collected data for analysis required several pre-processing steps to ensure that they were in a suitable form.

The raw data collected from different sources required pre-processing to make them suitable for analysis. Several important steps were taken to prepare the data for aspect-oriented sentiment analysis.

a. Tokenization:

In the first stage of data preprocessing, tokenization into sentences was performed. The feedback was divided into separate sentences, which will provide a more accurate and detailed analysis.

b. Extracting aspectual phrases:

Then, aspect phrases are extracted, which are certain phrases or expressions that convey information about certain aspects or characteristics of the smartphone under discussion. For example, in the sentence "This phone has a bad camera", the aspect phrase would be "bad camera".

The process of aspect phrase extraction included the following steps:

1) English translation: to ensure compatibility of Russian-language reviews with English-language tools, we translated tokenized sentences from Russian to English using Google Translate API.

2) Identification of aspect phrases. In order to remove unnecessary words and identify aspect phrases, the English translated sentences were syntactically analyzed using spaCy parser, which proved to be very effective for English texts. This step involved identifying syntactic links and relations in the text in order to eliminate unnecessary words and highlight key phrases.

3) Reverse translation into Russian: Aspect phrases received in English were translated back into Russian using Google Translate API.

4) Word2Vec synonym substitution: In order to preserve the semantic integrity of the original Russian-language reviews, Word2Vec was used to find synonyms of the translated words [9]. This step aims at preserving the semantic integrity of the original Russian-language reviews, since in the process of translation from English into Russian, the translator may replace some words with synonyms, which may affect the quality of aspect detection in aspect phrases later on.

Word2Vec compared each translated word with the original Russian word, and if a synonym with high cosine similarity is found, it replaces the word in the translated aspectual phrase. This approach ensures cohesion and thematic consistency of the phrases throughout the analysis.

c. Data cleaning and transformation:

The last stage of preprocessing is to clean the aspect phrases from inconsistencies and random "noise". This process includes removing special characters, empty strings and values, duplicates and stop words from the dataset.

Results of preliminary data processing

The results of data preprocessing are plotted in Table 1, which shows the number of documents in the dataset and the size of the dataset before and after preprocessing.

Table 1. Pre-processing results of the collected unstructured raw data

Data Set	Number of documents	Data set size
Before pretreatment	478 411	319,708 KB
After pretreatment	604 851	479 249 KB

The pre-processing results of the collected raw data showed a significant increase in the number of documents from 478,411 to 604,851 documents. Splitting the feedback into sentences and extracting aspect phrases increased the number of documents in the raw dataset to 604,851 documents, as each aspect phrase became a separate document in this dataset. The size of the dataset also increased from 319,708 KB to 479,249 KB.

After completing the data preprocessing step, the next important step in the study is to extract aspects from the aspect phrases obtained earlier. For this purpose, topic modeling techniques including Latent Dirichlet Allocation (LDA), Gibbs sampled Dirichlet polynomial mixture (GSDMM) and BERTopic are used. These methods allow us to identify and categorize the main themes or aspects present in smartphone reviews.

The most popular topic modeling approach is Latent Dirichlet Distribution (LDA), which is a generative probabilistic model algorithm that reveals latent variables representing abstract

topics to control document semantics [10]. LDA assumes that each document is a mixture of different topics, and each topic represents a probability distribution over a set of words.

GSDMM is another topic modeling approach that uses Gibbs sampling for faster convergence of the extracted topics. GSDMM is an extension of the Latent Dirichlet Distribution Algorithm (LDA) and is specifically designed for topic detection in small documents, assuming there is only one topic in the document [11]. GSDMM can be particularly useful for textual data where the number of clusters or topics is not known in advance, as it automatically determines the number of topics from the data. The algorithm of GSDMM is presented in Figure 2.

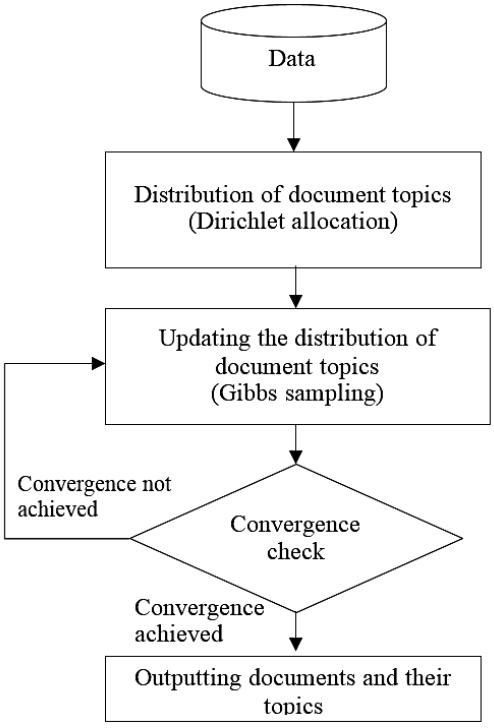

Figure 2. Algorithm of GSDMM operation

GSDMM uses the latent Dirichlet distribution and Gibbs sampling in its work.

First, each input document is assigned an initial topic based on a Dirichlet distribution. Gibbs sampling is used to update, at each iteration of the document traversal, the topic assignments for the documents in the corpus based on document words and topic words. The document selection at each iteration is randomized and its topic assignment is updated based on the words in the document and the current topic. This process is repeated for a given number of iterations until the topic and word assignments converge to a stable solution, stable topic convergence.

BERTopic is a topic modeling technique that uses transformer-based language models such as BERT (Bidirectional Encoder Representations from Transformers) for efficient and accurate topic modeling. BERTopic identifies clusters of similar documents and categorizes them into topics based on their semantic similarity [12].

BERTopic is a variant of the Latent Dirichlet Distribution Algorithm (LDA) topic modeling algorithm that identifies clusters of similar documents and assigns them to topics.

The stages of BERTopic's operation are summarized in Figure 3.

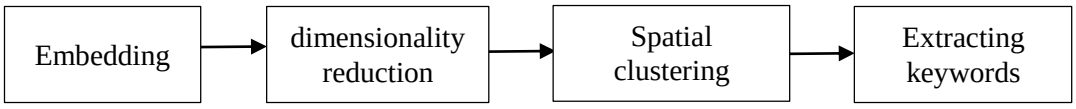

Figure 3. Stages of BERTopic operation

The operation of the algorithm can be visualized as 4 steps:

1. Embedding. This step converts text documents into numerical vectors called embeddings. Each embedding represents the value of a text document and is used to compare similarities between different documents.

2. Dimensionality reduction. To reduce the dimensionality of vectors and computational complexity, UMAP (uniform manifold approximation and projection) method is used, which preserves the local and global structure of multidimensional data by allowing similar documents to be grouped together.

3. Spatial clustering. It involves clustering attachments into groups (topics) based on their similarities. For clustering, BERTopic uses HDBSCAN spatial clustering, which identifies clusters based on the density of attachments.

4. Extracting keywords. Allows to extract the most important words for each topic using the c-TF-IDF method.

Three topic models, LDA, Dirichlet Multinomial Mixture model-based approach for short text clustering (GSDMM) and BERTopic, were used to implement the above topic modeling methods [13, 14].

After the aspect extraction step, the next important step in aspect-based tone analysis is tone analysis. The aim is to identify the sentiment polarity (positive or negative) associated with each aspect discussed in smartphone reviews. For this purpose, this study uses three different models for tone analysis: the SVM, LSTM and BERT support vector method.

SVM is a popular supervised machine learning method used for tone analysis problems [15]. SVM is effective for linear and nonlinear problems. SVM is particularly effective for binary classification tasks, where the goal is to separate the data into two classes or groups.

The basic idea of SVM is to find the optimal hyperplane that best separates data points belonging to different classes in the feature space. A hyperplane is a decision boundary that separates data points belonging to different classes in the feature space.

LSTM (Long Short Term Memory) is one of the most common types of recurrent neural network (RNN), which is commonly used in natural language processing tasks [16].

Unlike traditional feed-forward neural networks, which process data in a unidirectional stream, LSTM has the ability to store and utilize information from previous time steps, allowing it to capture long-term dependencies in the data. This ability allows LSTM to capture the context and emotional nuances of language, making it a powerful tool for analyzing tone.

BERT is a state-of-the-art natural language processing method that utilizes transformer-based language models [17].

BERT uses a bidirectional approach, considering both the left and right contexts of words in a sentence, allowing it to capture complex linguistic nuances and relationships between words. By understanding contextual relationships, BERT can capture subtle nuances, enhancing its ability to effectively discern tone in text.

Implementation of runtime tone analysis methods using SVC [18], LSTM [19] and RuBERT-tiny [20] models.

Results and discussion. This study compared the performance of three topic modeling methods for extracting aspects from smartphone reviews in combination with a cross-lingual parser. Each model was trained on a subset of randomly selected 100,000 reviews from the dataset and evaluated on four criteria: number of extracted smartphone aspects, coherence score,

algorithm training time, and aspect detection accuracy on test data.

Table 2 shows the comparison results of different topic modeling techniques for aspect extraction combined with a cross-lingual parser. Evaluation metrics include the number of identified smartphone aspects, consistency score, training time, and aspect identification accuracy on the test dataset.

Also for comparison, Table 3 shows the results of the comparison of thematic models, but using the Russian-language parser instead of the cross-lingual one.

Table 2. Results of comparison of thematic models for aspect detection in combination with a cross-lingual parser

Model	Highlighted aspects of smartphones	Consistency assessment	Algorithm training time (min)	Accuracy of aspect detection on test data (%)
1. LDA	7	0.38	18	56.7
2. GSDMM	8	0.46	32	61.9
3. BERTopic	8	0.55	43	73.5

The results obtained in Table 2 show that GSDMM and BERTopic identified the highest number of aspects, identifying eight aspects. They are followed by LDA with seven aspects.

Consistency score [21] is a metric used to assess the quality of the topics extracted by the models. A higher consistency score indicates more reliable and stable results. BERTopic showed the highest consistency score of 0.55, indicating more reliable aspect identification compared to LDA (0.38) and GSDMM (0.46).

In terms of training time, BERTopic required the longest training time: 43 minutes. GSDMM had a training time of 32 minutes, while LDA was the fastest, completing training in 18 minutes. Although LDA was the fastest, it had the lowest consistency score, which is more important than the training speed.

BERTopic achieved the highest aspect detection accuracy on the test data, scoring 73.5%. GSDMM followed with 61.9% accuracy, while LDA showed the lowest accuracy at 56.7%.

Table 3. Results of comparison of thematic models for aspect detection in combination with the Russian-language parser

Model	Highlighted aspects of smartphones	Consistency assessment	Algorithm training time (min)	Accuracy of aspect detection on test data (%)
1. LDA	7	0.38	17	56.4
2. GSDMM	8	0.47	35	61.8
3. BERTopic	8	0.56	47	72.2

Next, let's analyze and compare the results from both tables:

Comparing the results, it can be said that they are almost the same between the two tables: thematic patterns in combination using cross-lingual parsing (results are shown in Table 2) and thematic patterns in combination using Russian-language parser (results are shown in Table 3).

All topic models identified the same number of aspects: the LDA identified 7 aspects, while GSDMM and BERTopic identified 8 aspects each.

In Table 1, the BERTopic model achieved the highest consistency score of 0.55, indicating that the aspects extracted by BERTopic were more consistent compared to LDA (0.38) and GSDMM (0.46).

In Table 2, the results are similar, with BERTopic again receiving the highest consistency

score of 0.56, and LDA and GSDMM receiving 0.38 and 0.47, respectively.

In terms of training time, LDA is the fastest to train in two tables. On the other hand, BERTopic, being a more complex model, requires longer training time compared to GSDMM and LDA.

In terms of aspect detection accuracy on the test data, BERTopic again demonstrates the highest accuracy among the three models in both tables. BERTopic's high performance in accurate aspect detection signifies its ability to better understand the nuances of language in smartphone reviews.

Having analyzed the results of aspect extraction using different topic modeling models, we move to the tone analysis phase. In the next stage of our study, we evaluate the effectiveness of different sentiment analysis models for classifying the sentiments expressed in smartphone reviews.

Table 4 shows the comparison results of three sentiment analysis models, SVC, LSTM and RuBERT-tiny for sentiment classification. The metrics used to evaluate the models are accuracy, completeness and F1 score, and the table is divided into two classes representing negative and positive sentiments. The best score values are highlighted in bold.

Table 4. Results of quality comparison of models for tone analysis

Class	Model	Precision	Recall	F1-Score
0	1. SVC	0.77	0.77	0.77
	2. LSTM	0.83	0.85	0.84
	3. RuBERT-tiny	0.83	0.90	0.86
1	1. SVC	0.76	0.76	0.76
	2. LSTM	0.85	0.83	0.84
	3. RuBERT-tiny	0.89	0.82	0.85

For Class 0:

Among the three models, the LSTM and RuBERT-tiny models achieved higher accuracy (0.83 and 0.83, respectively) compared to the SVM model (0.77).

The RuBERT-tiny model achieved the highest completeness value (Recall) of 0.90, followed by the LSTM model (0.85), and the SVM model had a completeness of 0.77.

In terms of F1-score, the RuBERT-tiny model had the highest score (0.86), followed by the LSTM model (0.84). The SVM model had an F1-score of 0.77.

For Class 1:

The RuBERT-tiny model achieved the highest accuracy (0.89), followed by the LSTM model (0.85), and the SVM model had an accuracy of 0.76.

The LSTM model achieved the highest completeness (0.83), followed by the SVM model (0.76), and the RuBERT-tiny model had a completeness of 0.82.

In terms of F1 score, the RuBERT-tiny model had the highest score (0.85), followed by the LSTM model (0.84). The SVM model had an F1 score of 0.76.

Overall, the RuBERT-tiny model shows the most balanced performance with high completeness and accuracy for both sentiment classes.

Conclusion. In conclusion, the results of the study showed that the combination of cross-language parsing with BERTopic outperforms Russian-language parsing with BERTopic, achieving higher accuracy on the test data. This emphasizes the effectiveness of this combination and the enhanced context understanding capabilities of BERTopic for more accurate aspect identification.

In addition, the study compared the SVC, LSTM, and RuBERT-tiny models for sentiment analysis, with RuBERT-tiny achieving the highest accuracy. These results make a valuable contribution to the selection of suitable methods for solving the problem of aspect-based tone analysis.

Overall, this study emphasizes the importance of selecting appropriate methods and approaches for aspect extraction and sentiment analysis. The combination of cross-lingual syntactic analysis with BERTopic is found to be more effective than other models in aspect extraction, and RuBERT-tiny shows high accuracy in the tone analysis task.

References

1. Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55, 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1
2. M. Gupta, A. Mishra, G. Manral and G. Ansari, "Aspect-category based Sentiment Analysis on Dynamic Reviews," 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020. – Pp. 492-496, doi: 10.1109/ICCCA49541.2020.9250914.
3. Hetal Gandhi, Vahida Attar. Extracting Aspect Terms using CRF and Bi-LSTM Model // Procedia Computer Science – 2020. – Volume 167. – S. 2486-2495. – ISSN 1877-0509. – doi: 10.1016/j.procs.2020.03.301.
4. M.A. Qureshi et al., "Sentiment Analysis of Reviews in Natural Language: Roman Urdu as a Case Study," in IEEE Access, vol. 10, pp. 24945-24954, 2022, doi: 10.1109/ACCESS.2022.3150172.
5. K. Maharana, S. Mondal, B. Nemade. A review: Data pre-processing and data augmentation techniques // Global Transitions Proceedings. Volume 3, Issue 1, June, pp. 91-99, 2022, doi: https://doi.org/10.1016/j.gltp.2022.04.020.
6. Sumit Kumar, Uponika Barman Roy. A technique of data collection: web scraping with python. - Statistical Modeling in Machine Learning, Academic Press. – 2023. – S.23-36. – doi: 10.1016/B978-0-323-91776-6.00011-7.
7. H. Hassani et al., "LVTIA: A new method for keyphrase extraction from scientific video lectures" in Information Processing & Management,vol. 59, Issue 2, 2022, ISSN 0306-4573, doi: https://doi.org/10.1016/j.ipm.2021.102802.
8. Huwail J. Alantari et al., "An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews" in International Journal of Research in Marketing, vol. 39, Issue 1, 2022, pp. 1-19, ISSN 0167-8116, doi: https://doi.org/10.1016/j.ijresmar.2021.10.011.
9. Goldberg, Y., & Levy, O. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. – 2014. – doi: 10.48550/arXiv.1402.3722.
10. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation - Journal of Machine Learning Research 3 - 2001. – S. 601-608.
11. Jianhua Yin, Jianyong Wang. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering - Association for Computing Machinery, New York, NY, USA - 2014. – S. 233-242. – doi: 10.1145/2623330.2623715.
12. Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. – 2022. - https://arxiv.org/abs/2203.05794.
13. C. Sharma et al., "Latent DIRICHLET allocation (LDA) based information modelling on BLOCKCHAIN technology: a review of trends and research patterns used in integration" in Multimedia Tools and Applications, vol. 81, Issue 25, pp. 36805-36831, 2022, doi: 10.1007/s11042-022-13500-z
14. A. Udupa, K. N. Adarsh, A. Aravinda, N. H. Godihal and N. Kayarvizhy, "An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling," 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP), Bengaluru, India, 2022, pp. 1-9, doi: 10.1109/CCIP57447.2022.10058687.
15. Bruno Stecanella. Support Vector Machines (SVM) Algorithm Explained. – 2017. – https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm.
16. Hochreiter, Sepp & Schmidhuber, Jürgen. Long short-term memory. – Neural computation - 1997. – No.9. – S.1735-80. – doi: 10.1162/neco.1997.9.8.1735.
17. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - 2018. - https://arxiv.org/abs/1810.04805.
18. scikit-learn. sklearn.svm.SVC. - https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

19. TensorFlow. TensorFlow v2.12.0 tf.keras.layers.LSTM. – https://www.tensorflow.org/api_docs /python/tf/keras/layers/LSTM.
20. HuggingFace. RuBERT-tiny. - https://huggingface.co/cointegrated/rubert-tiny.
21. M. R¨oder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures // In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. – 2015. – S.399-408. – doi:10.1145/2684822.2685324.