Sentiment Analysis of Information Space as Feedback of Target Audience for Regional E-Business Support in Ukraine

Victoria Vysotska [1,2], Oksana Markiv [1], Stepan Tchynetskyi [1], Bohdan Polishchuk [1], Oksana Bratasyuk [2,3] and Valentyna Panasyuk [3]

[1] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Osnabrück University, Friedrich-Janssen-Str. 1, Osnabrück, 49076, Germany*
[3] *West Ukrainian National University, Lvivska Street, 11, Ternopil, 46004, Ukraine*

Abstract

In conditions of the war in Ukraine, e-business plays a key role in supporting and developing the economy of country, maintaining business relations and competitiveness in the international area of the financial market, interacting with government bodies and supporting feedback from the target audience. The paper describes the application of sentiment analysis of comments, feedback, requests and news for the support and development of e-business. The analyzed analogs made it possible to develop information technology for solving NLP problems of e-business, adapted for the Ukrainian target audience. The general typical structure of the information system for the support and development of e-commerce has been developed by analyzing the feedback of the target audience based on machine learning technology and natural language processing methods. The logistic regression method coped best with the task of analyzing the impact of the news on the financial market, which has shown an accuracy of 75.67%. This is certainly not the desired result, but it is the largest indicator of all the considered. The support vector method (SVM) has shown an accuracy of 72.78%, which is a slightly worse result than the one obtained with the help of the logistic regression method. And the naive Bayesian classifier method has shown the worst accuracy of 71.13%, which is less than the two previous methods.

Keywords 1

Sentiment analysis, feedback, comment, e-commerce, e-business, NLP, machine learning, content analysis, personal data, information security, personal data protection

1. Introduction

Business plays a key role in the economy of every country. Thus, in Ukraine, small and medium-sized businesses provide about 64% of added value, 81.5% of employed workers in economic entities and 37% of tax revenues in 2021 [1]. Due to the war in Ukraine, a large part of small and medium-sized businesses has been liquidated (especially in the occupied territories), or has moved, or has switched completely into the field of e-commerce. A big problem concerning e-businesses is that there is not enough information about development opportunities in certain locations and no feedback from their consumers or such information comes late or incomplete, or with excessive noise. In the conditions of war, it is also worth talking not only about the development of e-business, but also about its recovery, because many enterprises stop their work or are completely destroyed in connection with the war. In such conditions, additional tools and information technologies are needed to help businessmen monitoring e-business development opportunities in a certain location, as well as establish feedback

MoMLeT+DS 2023: 5th International Workshop on Modern Machine Learning Technologies and Data Science, June 3, 2023, Lviv, Ukraine
EMAIL: victoria.a.vysotska@lpnu.ua (V. Vysotska); oksana.o.markiv@lpnu.ua (O. Markiv); stepan.tchynetskyi.msaad.2022@lpnu.ua (S. Tchynetskyi); Bohdan.Polishchuk.SA.2018@lpnu.ua (B. Polishchuk); rosoliak@gmail.com (O. Bratasyuk); v.panasiuk@tneu.edu.ua (V. Panasyuk)
ORCID: 0000-0001-6417-3689 (V. Vysotska); 0000-0002-1691-1357 (O. Markiv); 0000-0002-5110-9423 (S. Tchynetskyi); 0000-0002-0545-6264 (B. Polishchuk); 0000-0002-5871-4386 (O. Bratasyuk); 0000-0002-5133-6431 (V. Panasyuk)

with users using social networks and mass media. Such tools will help significantly to expand the vision of market opportunities for e-business and will clarify which of them make sense to invest in. And finally, to see what idea the future holds and what business model needs to be implemented/ maintained/developed for the rapid development of territorial/interregional e-business. It will also help to understand which levers have the greatest effect for changing business policy: what should stay the same, and what to change to ensure high speed in the implementation of the plan based on the analysis of relevant research results, for example, to receive:

- Direct feedback from customers, dynamics of changes in overall satisfaction or interest of the target audience and advantages/disadvantages from users using NLP analysis.
- Support for the development of e-business in relation to the location of their enterprise and the best directions of development.
- Schedules of business development (improvement/deterioration) depending on the content of the comments.

Building a serious and thriving e-business in any customer-facing industry requires time and attention to serve those customers. After all, customer service teams interact directly with potential customers every day [2]. It can bring both the greatest benefit and the greatest loss. When customer service is a priority, companies receive many benefits: more loyal customers, more positive reviews, and much more revenue. That's why it's so important to be focused on customer service. Providing customer support can take a lot of time and energy, so traditional customer service is often seen as a cost center. Business leaders know they need to provide services, but they see it as a "cost of doing business." However, communicating with customers can be just as profitable as developing the product itself. Customer service is not the only cost of doing business. It is an important part of the overall customer experience. However, good customer support can lead to high costs, which is never a good thing, especially for smaller companies or those just starting their commercial journey. That is why more and more companies [3] are starting to transfer the problems of organizing and maintaining a good and efficient service center to other outsourcing companies or startups.

Therefore, it is relevant to analyze the directions of building information technology to support the development of e-business in Ukraine by analyzing business locations, processing feedback from users, analyzing and classifying customer reviews in real time from social networks: Twitter, Reddit, Facebook and others using methods of deep learning and Natural Language Processing of Ukrainian and English-language texts.

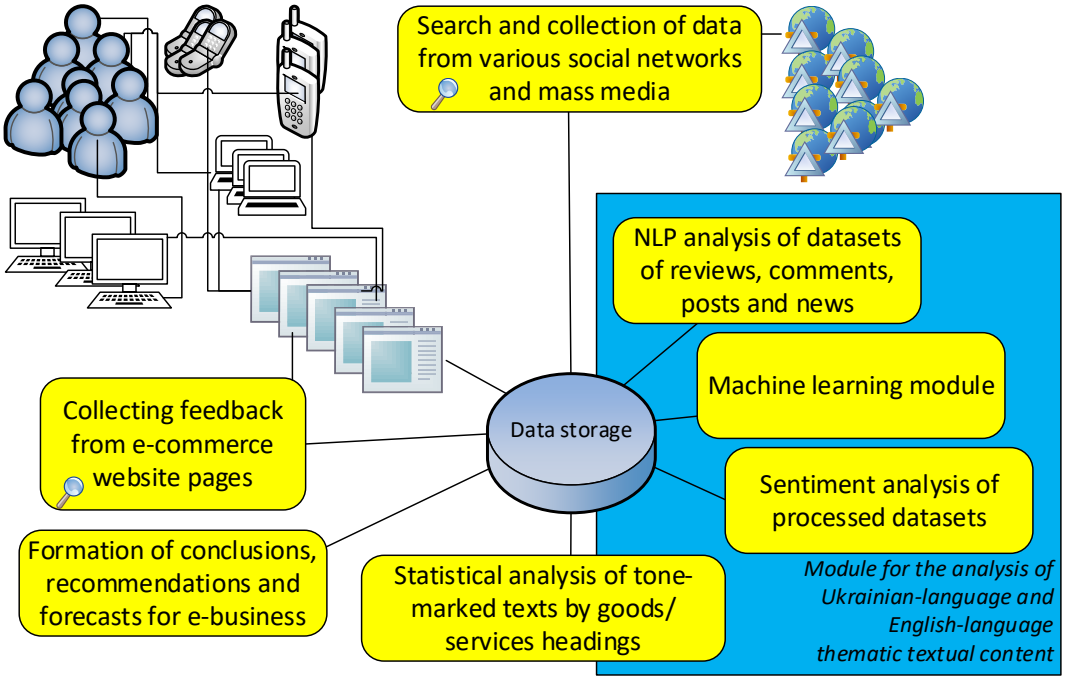


Figure 1: General scheme of the information space sentiment analysis

When analyzing reactions to goods and services through the analysis of comments, feedback on them on websites, in social network profiles and in parallel with news in the public about similar goods and services, etc. the input data comprises Ukrainian-language and English-language content from the specific e-business own sites, from the social network profiles of regular customers and the profile of the company itself, and in parallel from reliable media sources, where possible news about these similar goods and services, for example, construction, etc. It is necessary to develop an approach for analyzing the backlash of the target audience for Ukrainian e-business, because in modern conditions it is e-commerce that survives more on the territory of Ukraine. So, there is a need quickly and efficiently automatically to collect and analyze the reaction of the target audience for the opportunity to direct business. In times of war, constant power outages, including non-scheduled light, business must adapt quickly without using tools and techniques standard for peacetime, including data collection, for example, to predict what will be more relevant and better implemented and for which audience (age, gender, region, etc.). It should be a technology for processing already collected data from reliable sources to extract certain reactions (sentiment analysis, the tonality of positive, neutral or negative feedback on a product, for example).

In general, users are often either illiterate, or accidentally write with mistakes, or use mixed language depending on the user's region, including English words inserted in English or transliterated but still with errors. The same users, especially young people, often write reviews in English on social networks. That is why three is combination in two languages and from different sources. So, it is similar to how a bot collects data from reliable sources and filters, and then forms a dataset (this is not described in detail in the article, because there are many similar publications, including authors). The article focuses only on the process of processing datasets in two languages based on NLP and MN. And the emphasis is more on sentiment analysis, which is to extract a primitive emotion in the text for a specific product or product, or a type of product or service from, for example, specific users so that there is a possibility further to analyze and forecast based on recommendations and collected statistics, for example, the general reaction to a category of goods from a certain class of the target audience. It is necessary to develop such a system, which is designed to simplify communication between customers and companies, especially for those companies that cannot afford a full-fledged support center. The peculiarity of this system will be in the use of NLP algorithms to reduce customer service costs by reducing the number of active employees in the company. Human power will be replaced by an artificial intelligence algorithm that will itself classify customer reviews and complaints and determine the necessary actions for them.

The purpose of the research is to develop information technology for the analysis of Ukrainian- and English-language user-client reviews on e-commerce sites, posts and news in social networks and mass media based on natural language processing methods and machine learning technology for the promotion, adaptation and further development of the relevant e-business.

To achieve the goal, the following tasks must be solved:

1. Research and comparison of analogues;
2. Comparison and research of modern NLP methods such as lemmatization and stemming, keyword extraction, sentiment analysis, text summarization, bag of words and tokenization;
3. Develop a model of the classification system of customer reviews and news from reliable sources to identify the emotional coloring of the text in Ukrainian-English based on the Naive Bayes classifier;
4. Carry out an experimental test of the developed sentiment analysis system of the information space as feedback from the target audience to support e-business in Ukraine.

2. Related works

The interaction between a company and its target audience has been studied for centuries. From the very beginning of commercial relations, the relationship between the service provider and the recipient has been valued almost above all else. Trade is built on trust and respect. The image of an entrepreneur is often more important than the product. For many hundreds of years, the relationship between the merchant and the buyer, the entrepreneur and the client has not lost its importance, and in the era of mass digitalization, the quality of the relationship between the company and the target audience of

different sizes and the professional support of customer feedback often determine the success of e-business [1].

The interaction between companies and customers is a complex relationship that needs to be maintained in a good way for companies. Every company should now have its own customer support center. However, such centers are expensive and, in the times of startups and companies that appear and disappear equally quickly, it is not profitable to create a home-based customer support center with hired staff. Now, in the time of global digitalization and even greater acceleration of the movement of life, it is unprofitable to have customer support centers that operate on the basis of agents. After all, the speed of business is increasing, and the number of new customers is also increasing with it. However, more customers are not only an increase in profits.

On the other hand, nowadays social networks occupy a large, perhaps even too large, place in the life of an average modern person, a potential client of a particular e-business. The speed with which news can spread across social networks is fascinating and frightening at the same time. And it is in this environment that companies have to communicate with customers. The price of poor customer service, including support, can be too high. That is why it is important to have a high-quality, efficient customer support center. It is the customer support centers that often determine the attitude of the general public towards the company. The company's relationship with the target audience increases not only e-business retention, but also serves as free advertising: if the customer likes the product and service, it is more likely to recommend own business to others or leave comment/feedback on a social network.

Customer support is one of the most important aspects of many enterprises and companies. However, it is not so easy. An effective customer support center requires a lot of expenses: agent salaries, their workplaces, agent training. These are all expenses. And for many companies, these costs are becoming too high. More and more companies prefer intermediary firms that specialize in communicating with the target audience of a particular e-business. It also requires certain costs and time for cooperation and training of personnel for specific e-business. In the modern age of digitization, it is the replacement of such call centers and intermediary firms with a tool in the form of an information system for interacting with customers and analyzing comments and news based on machine learning and NLP methods that can become a successful business solution. NLP allows to apply machine learning algorithms to text and speech. For example, it is possible to use NLP to create systems such as speech recognition, document summarization, machine translation, spam detection, named entity recognition, question answering, autocompletion, predictive text input, etc. [4]. Thanks to the latest and/or classical algorithms, for example, the Turing test [5], the system can compete with the leading companies in the outsourcing market and, potentially, change the rules of customer interaction. Then even small companies will be able to easily maintain only a few agents, but have the same quality of support as the giants of their industry with multiple budgets, for example, based on modeling, synthesis and speech recognition technology [6].

Now there is also a very relevant problem of solving NLP problems for Slavic languages, especially the Ukrainian language against the background of the war in Ukraine (for example, for identifying fakes and propaganda, it is even relevant for e-business - an example will be whether or not the war in Taiwan changes the price policy on all digital devices), which would allow Slavic countries to qualitatively use such NLP solutions as: text generation; sentiment analysis; generalization of the text; and other.

Outsourcing is a company strategic decision to reduce costs and increase business efficiency by hiring an individual/legal entity to perform relevant tasks [7]. Outsourcing customer support is a fairly common practice (for example, Sykes [8], Sensee [9], Serco [10], Teleperformance [11]), so the market of outsourcing companies specializing in communication with customers is quite extensive (Table 1). It is possible to find a solution for almost any e-business. However, if to create a startup as an analogue of performing at least part of the tasks of the relevant outsourcing companies, which will be more economical or more efficient, then this will greatly undermine the already established market. After analyzing the various companies and the services they offer, as well as their pros and cons, a set of characteristics and evaluation criteria for a customer interaction system was developed:

- 24/7 support access - the presence or absence of 24/7 communication support is assessed;
- Speed of feedback - how many hours on average between all channels are required to provide the first response to the client;
- Confidentiality;

- Number of agents - the value of the number of agents should not be too high and not too low;
- Location and size of the office - the location of the office should allow reaching the largest number of clients, the size of the office should provide a workplace for all the company's agents;
- Number of available communication channels - voice, text, chat;
- Possibilities of inbound/outbound communication, telemarketing, active feedback collection;
- Price and number of languages.

Table 1
Well-known customer support outsourcing companies

Title	Advantages	Disadvantages
Sykes [8]	Agents from several geographical locations (EU, UK and USA); Focused on the holistic path of the client; Compliant with HIPAA and certified by PCI for working with sensitive data; Multi-channel communication; Provides strategic advice; Flexible and scalable.	Mainly voice communication channels (70%).
Sensee [9]	Ethical customer support service; Work around the clock; ISO-accredited; Focus on a single brand.	Potentially a smart choice only for small businesses and those in the financial industry (e.g. credit card companies).
Serco [10]	24/7 support; Diverse workforce; Processes confidential and secure data.	Human-oriented with a slight emphasis on technology. Good for public sector only.
Teleperformance [11]	Diverse workforce; 265 languages and dialects; Great language skills for companies with a global client base; Multi-channel communication; Focus on analytics; Provides multi-channel support.	Not very suitable for companies looking for a more personalized approach.

Another area of information and sentiment gathering that affects the development of e-business of a certain sector is the monitoring platforms of global/regional media and print media, social, online, digital and broadcasting companies such as Carma Media Monitoring, Repustate. Patient Voice [12-13], Siri [14], Grammarly [15], Klevu Smart Search [16], etc. (Table 2).

Table 2
Well-known analytical data collection tools based on NLP and machine learning methods

Title	Advantages	Disadvantages
Carma Media Monitoring	Ability to monitor feedback online; Impressive dashboards; Online support;	The cost of the platform; Difficulty of use; It is more suitable for well-known companies.

	Support for review of social networks, newspapers and television.	
Repustate. Patient Voice [12-13]	Ability to receive quality feedback from customers; Analysis of social networks; A large number of NLP methods.	The cost of the platform; The company must be big.
Siri [14]	Ease of use; Recognition of timbres; Ease of use; Great functionality.	Limitation in the use of languages.
Grammarly [15]	Ease of use; Evaluation of the text; Selection of text style.	English only; Incorrect offers; Ignorance of tone and context; Suppression of writers' freedom of speech.
Klevu Smart Search [16]	Ease of use; Support for usage analytics; 24/7 support.	Inaccuracies in predictions; Connection cost.

Usually, products that use NLP in business are very convenient, but limited functionality does not allow users to fully cover their needs. Therefore, in the developed product, it is necessary to attract all the advantages of analog products, expand the functionality of the product, which would cover all the needs of customers and basically correct the shortcomings of analog products. The best analogue is Repustate, it should be the main competitor to be bypassed. This product involves a large number of NLP techniques, as expected in the product under development. All the other products discussed above are made using NLP techniques and are leaders in their fields, so having their expertise can involve their approaches as an extension of functionality for the product being developed, making it a market leader in products that involve NLP.

Special attention should be paid to the security of personal data of customers. If even when they write negative reviews and want to remain relatively anonymous to the general readership, they have the right to do so. Any e-business must take into account all customer opinions, not only positive ones, in order to successfully direct its business policy and quickly respond to certain shortcomings. Trust between the client and the business is built on trust and quality of service. Therefore, to maintain a high level of trust, the main point is to observe the security of personal data.

One of the negative consequences of the introduction of information and telecommunication technologies in all spheres of public life is the violation of important human rights, which manifests itself in the illegal collection, use and dissemination of personal data, including on the Internet. Inadequate legislative protection and insufficient protection of personal data in this area have led to an increase in human rights violations. Respect for the right to privacy is the basis of social justice and harmony. One of the most problematic legal aspects in the information technology era is the protection and security of personal data.

The Civil Code of Ukraine provides that an individual has the right to freely collect, store, use and disseminate information. It is not allowed to collect, store, use and disseminate information about the personal life of an individual without his or her consent, except in cases specified by law and only in the interests of national security, economic well-being and human rights. A person who disseminates information is obliged to make sure that it is accurate. A person who disseminates information obtained from official sources (information of state authorities, local self-government bodies, reports, transcripts, etc.) is not obliged to verify its accuracy and shall not be liable in case of its refutation. A person who disseminates information obtained from official sources is obliged to make a reference to such a source.

According to the Law of Ukraine "On Personal Data Protection", personal data means information or a set of information about an individual who is identified or can be specifically identified; personal data subject means an individual whose personal data is processed. The consent of the personal data

subject should be understood as a voluntary expression of the individual's will (subject to his or her awareness) to grant permission to process his or her personal data in accordance with the specified purpose of their processing, expressed in writing or in a form that allows to conclude that it has been granted. In the field of e-commerce, the consent of a personal data subject may be provided during registration in the information and telecommunications system of an e-commerce entity by marking a note on consent to the processing of his or her personal data in accordance with the specified purpose of their processing, provided that such a system does not create opportunities for processing personal data before marking.

- Personal data owner - an individual or legal entity that determines the purpose of personal data processing, establishes the composition of this data and procedures for its processing, unless otherwise provided by law.
- Personal data manager - an individual or legal entity authorized by the personal data owner or by law to process this data on behalf of the owner.
- The use of personal data is any actions of the owner to process this data, actions to protect it, as well as actions to grant partial or full rights to process personal data to other subjects of relations related to personal data, which are carried out with the consent of the personal data subject or in accordance with the law.
- The personal data owner is obliged to use personal data if he/she creates conditions for the protection of this data. The controller is prohibited from disclosing information about the personal data subjects whose personal data has been provided to other parties to the relations related to such data.
- The dissemination of personal data involves actions to transfer information about an individual with the consent of the personal data subject.
- Dissemination of personal data without the consent of the personal data subject or his/her authorized person is allowed in cases specified by law and only (if necessary) in the interests of national security, economic well-being and human rights.

Owners, managers of personal data and third parties are obliged to ensure the protection of this data from accidental loss or destruction, unlawful processing, including unlawful destruction or access to personal data. State authorities, local self-government bodies, as well as owners or managers of personal data that process personal data subject to notification in accordance with this Law, shall establish (appoint) a structural unit or a responsible person to organize work related to the protection of personal data during their processing.

3. Materials and methods

NLP combines computational linguistics, rule-based modeling of human language, with statistical, machine learning, and deep learning models. Together, these technologies allow computers to process human speech in the form of text or voice data and "understand" its full meaning, taking into account the intentions and moods of the speaker or writer [17]. NLP has become an important business tool for uncovering hidden data from social media channels. Sentiment analysis can analyze the language used in social media posts, responses, reviews and more to extract attitudes and emotions in response to products, promotions and events – information that companies can use in product design, advertising campaigns and more.

It is also appropriate to use NLP to classify customer feedback. The only external action required to start the system is the client writing a review. This review can be written on any platform: from social networks to Google Maps. The specifics of the number and which platforms are agreed upon by the company using the system. After the customer has written his review, the system downloads this review from the specified platform to its own storage. In this way, a feedback bank is built, which can be used in further iterations of the system model (Fig. 1). When the feedback is collected and written to the repository, the system performs the feedback classification operation. This means that the system determines whether a new review is positive or negative, checks whether any action is required on that review, and which word from the review best describes the review in general. After successful classification, depending on the results, the system stores the feedback in another repository for archiving and forwards the information to agents if needed.

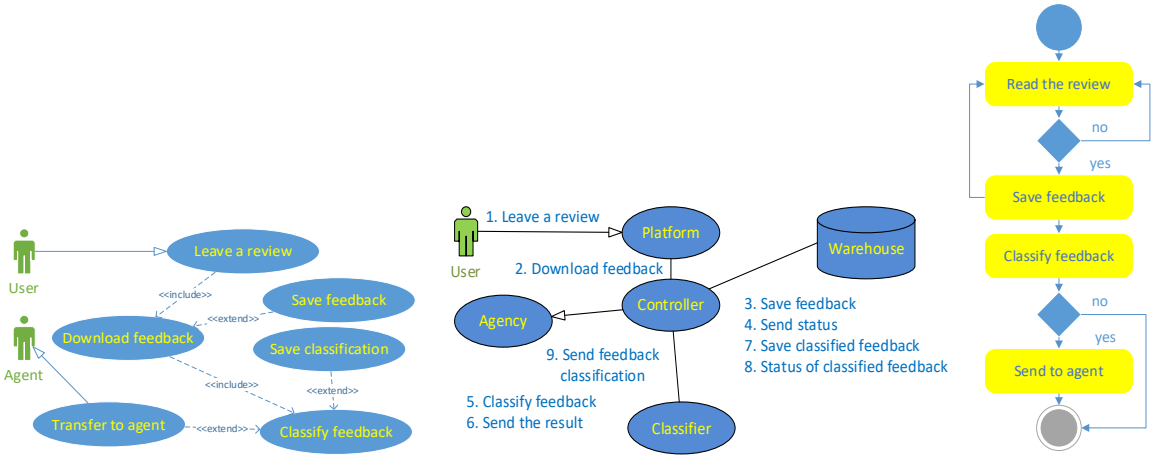


Figure 2: Use case, cooperation and activity diagrams

Since the resource is planned to be online, the devices available to the user will be used to interact with the user. When a user enters one of the designated platforms, he must click the appropriate button to leave a review. After the user has sent his feedback, the feedback is automatically collected by the system controller (Fig. 2).

The controller sends this feedback to the Storage, which stores the raw feedback. After the Vault has performed a save, it sends a response status back to the Controller for logging. Then, when the Controller has received a feedback message from the Repository, it sends the feedback to the Classifier. The classifier classifies the response. Then, the already classified feedback is sent back to the Classifier, which, in turn, sends information about the classified feedback to the Repository, so that it stores the feedback again, but already in a processed form. After saving the feedback, the Repository again sends the status of the saving to the System Controller, where it continues the flow, namely, sending the feedback to the Agency. The agency, depending on what the Classifier predicted, either forwards the feedback to the agents or terminates the feedback path.

The system constantly monitors available platforms for new reviews. The cycle of checking for new reviews continues until at least one new review is found on any platform. If a new response is found, the system breaks out of the cycle and starts active work (Fig. 2).

First, the new feedback is stored in the repository. The repository receives any feedback that has passed the previous stage, so it is possible that the same or similar in meaning and structure may be present in the repository. In any case, when new feedback comes into the system and it is recorded in the repository - the system passes the new feedback down the funnel and returns to monitoring new feedback. Thanks to this, new reviews will not accumulate, which is important for the speed of processing all reviews. After saving and passing on the feedback, the most costly action of the entire system takes place - classification. All the main calculations of the system take place here, which makes it a critical point for system efficiency. It is important to optimize this activity. After classification, depending on the results, the feedback is either passed to the agents for further action or sent to the repository for possible further use, such as analysis, archiving, improvement and iteration of the classification models. If the system decides that the feedback needs action, it sends it to the agents. Agents should resolve the issue raised by the feedback as soon as possible.

Human language is amazingly complex and diverse. People express themselves in endless ways, both verbally and in writing. Not only are there hundreds of languages and dialects, but each language has a unique set of grammatical and syntactic rules, terms, and slang. When people write, they often make mistakes, shorten words or miss punctuation marks. There are also regional accents, mumbling, stuttering, and borrowing terms from other languages, including Ukrainian [18].

All business data contains a lot of useful information, ideas, and NLP can quickly help companies get them. NLP tools process data in real-time, 24/7, and apply the same criteria to all data, so the results are accurate – and free of inconsistencies. Once NLP tools can understand what text is about, and even measure things like sentiment, companies can begin to prioritize and organize their data in ways that suit their needs [19].

Two main algorithms can be used to solve NLP problems: rule-based and machine learning. The biggest advantage of machine learning algorithms is their ability to learn on their own. There is no need to define rules manually – instead, machines learn from previous data to make predictions on their own, allowing for greater flexibility.

But before using machine learning methods, any text, either in English or in Ukrainian, or a mixture of them, may be pre-processed by NLP methods, in particular or partially depending on the purpose and type of task, taking into account the peculiarities of the method:

- Thematic analysis is extracting meaning from the text by identifying recurring themes based on machine learning [20];
    - Topic modeling can infer patterns and group similar expressions without having to define topic tags or train data beforehand;
    - Text classification or topic extraction from text must know the topics of the text before starting the analysis, as one needs to label the data to train the classifier.
- Sentiment analysis is determining whether a text is positive, negative or neutral based on other NLP and machine learning techniques to assign weighted sentiment scores to objects, topics, themes and categories in a sentence or phrase [21];
- Intent detection uses machine learning and NLP to automatically associate words/phrases with a specific intent. For example, a machine learning model can learn that the words buy or purchase are associated with purchase intent [22];
- Keyword extraction is a text analysis technique that automatically extracts the most used and most important words/expressions from the text [23-24];
- Lematization is grouping of different inflectional forms of a word for further analysis as a single element and, unlike stemming, brings context to words, i.e., connects words with similar meanings into one word; use positional arguments as input, for example, whether a word is an adjective, a noun, or a verb [25-26];
- Stemming is used to remove suffixes from words and ultimately obtain the so-called word base, which allows to standardize words to their base regardless of their inflections, for example for clustering or text classification and search [25-26];
- Tokenization is a method of dividing a text fragment into smaller units (tokens) and is used in traditional NLP methods (Count Vectorizer) and in architectures based on advanced deep learning (Transformers); markers can be words, symbols or sub-words (n-grams of symbols) [27];
- Machine translation is the task of automatically converting one natural language into another, saving the value of the input text and creating a free text in the output language [28-29];
- Generalization of the text is semantic reduction of the text by removing unimportant text and transforming the same text into a smaller semantic text form without removing the semantic structure of the text [30]; identifying important phrases in a document and using them to identify relevant information to add to the summary is a critical task for extraction-based summarization [31].

Table 3

Basic NLP methods for text classification

Title	Peculiarity	Advantages
Thematic analysis	Companies create and collect huge amounts of data every day. The tool will help businesses make better decisions, optimize internal processes, identify trends, and provide all kinds of other benefits to make them more efficient and productive [20].	Automated topic analysis with machine learning allows to scan as much data as someone want, providing new opportunities for meaningful insights; Real-time analysis - By combining topic detection with other types of natural language processing techniques, such as sentiment analysis, it is possible to get a real-time picture of what customers are saying about a product.

		Consistent criteria is a combination of statistics, computational linguistics, and computer science, so it is possible to expect high-quality results with unmatched accuracy.
Sentiment analysis	Helps data analysts in large enterprises to estimate public opinion, conduct detailed market research, monitor brand and product reputation, and understand customer experience. In addition, data analytics companies often integrate third-party sentiment analysis APIs into their own customer experience management systems, social media monitoring, or workforce analytics platform to provide useful insights to their customers [21].	Data Sorting at Scale - helps businesses process huge volumes of unstructured data in an efficient and cost-effective way; Real-time analysis – helps to identify customer dissatisfaction in real time, identify critical issues with feedback; Consistent criteria - Companies can apply the same criteria to all their data, helping them improve accuracy and gain better insights.
Detection of intentions	However, intent classifiers must first be trained with text examples, otherwise known as training data. Intent classification allows businesses to be more customer-centric, especially in areas such as customer support and sales. From responding more quickly to potential customers to dealing with high volumes of inquiries and offering personalized services, intent detection can be a key tool [22].	Use every sales opportunity - early detection of purchase intent is critical to sales and customer support, as it allows companies to take immediate action and convert leads into paying customers. Scale as you grow - Intent classifiers can pinpoint interested prospects and direct those specific inquiries to sales. Consistent criteria - ensures that all customer intents are analyzed under the same circumstances using the same standards, protocols, algorithms, etc. This significantly reduces errors and improves data accuracy. Increase Conversion Rate in Sales Campaigns - Identify high-intent leads and follow up with them immediately. Thus, conversion rates increase dramatically. Get analytics from shopping campaigns - help easily to create reports based on actual data about conversion rates, interested buyers, upsell opportunities and more.
Extracting keywords	It helps to summarize the content of the texts and to recognize the main topics that are discussed. Keyword extraction uses machine learning (AI) artificial intelligence with NLP to break down human language so that it can be understood and analyzed by machines. It is	Scalability - automatic keyword extraction allows to analyze as much data as is needed. It would be possible to read the texts and identify the key terms manually, but this would take a very long time. Automating this task gives the

	used to search for keywords in any text: ordinary documents and business reports, comments on social networks, online forums and reviews, news reports, etc. [23]. Organizations can automate some of the most routine tasks, saving valuable time and resources when analyzing data. Can be used to get valuable information about products/services to make decisions based on this data.	freedom to focus on other parts of your work. Consistent criteria - keyword extraction operates on the basis of rules and predefined parameters, there is no need to deal with inconsistencies that are common in manual text analysis. Real-time analysis - the ability to highlight keywords in social media posts, customer reviews, surveys or support requests in real-time, and get an idea of what is being said about the product, how it is happening and monitor them over time [24].

Stemming and Lematization are widely used in text analysis, where Text Mining is a method of natural language text analysis and extraction of high-quality information from the text (Table 4) [26].

Table 4

Comparison of Stemming and Lematization

№	Stemming	Lemmatization
1	Stemming is faster because it cuts words without knowing the context of the word in the given sentences	Lemmatization is slower, but it knows the context of the word before proceeding.
2	It is a rule-based approach	This is a dictionary-based approach
3	Less accurate	The accuracy is greater
4	When turning any word into its root form, stemming can create the meaning of a non-existent word.	Lemmatization always gives the dictionary meaning of a word when transformed into the root form
5	Stemming is used when the meaning of the word is not important for the analysis. Example: spam detection	Lemmatization would be recommended when the meaning of the word is important to the analysis Example: Question Answer
6	Example: word as information in Ukrainian (інформації [informatsiyi] =>інформац [informats])	Example: word as information in Ukrainian (інформації [informatsiyi] =>інформація [informatsiya])

To create a text classification module, it is firstly needed to determine which machine learning algorithm is best for purposes. There are many different classification algorithms, each with its own advantages and disadvantages. To begin with, it is proposed to determine the data with which to work, the power of the machine equipment and the optimal time for which the algorithm should produce the result of its work. Data collected from reviews on Google Maps is used to train the models. These data are the texts of Google Maps users' reviews of various establishments: restaurants, hotels, cafes, shops, etc. The dataset includes the feedback itself, recorded in the form of a tape, to which class this feedback belongs to the class of positive feedback, or to the class of negative feedback, as well as to which class this feedback belongs to the need for help / actions. In general, there are three indicators in the dataset. Reviews are written in Ukrainian, which significantly complicates the task. Also, in total, there are approximately 500 rows of data in the dataset. As part of the functions of this process, the Ukrainian dictionary of the GitHub user DICT_uk was used, which contains more than a million Ukrainian words, their meanings, belonging to parts of the language, and more [32]. As for power, prediction experiments will be performed on a local machine with an Intel Core i7-9750H 2.6 GHz and 12 GB processor. RAM.

To process this dataset, a machine with such power should be enough. In the worst case, the power of the machine used will affect the processing time of the data. For the prototype, this is not critical, but for the system itself, the hardware must be powerful enough to process feedback both individually and in queue mode one by one. Regarding the time for which the algorithm should issue information. The algorithm itself, for one response, provides results relatively quickly, up to half a minute. The most time-consuming part is the actual training of the model using the algorithm. During operation, the model will already be trained, and if it is necessary to retrain the model, it can be done during low system usage. However, future efforts should be made to optimize the model to reduce the feedback processing time to a minimum. Now that the conditions are defined, it is necessary to analyze and choose the best algorithm to work with. Among the most possible candidates, such classifiers as [33-34] are distinguished:

- Naive Bayes Classifier is a group of very simple classification algorithms based on Bayes theorem; all attributes of the dataset are independent and that none of them affects any other; is fast and requires little data for training, also has a good tendency to work with text problems, especially NLP;
- Support vector machines (SVM) is an algorithm used for classification and regression problems; divides the data into two half-planes with the best possible result, that is, finds such a line on the data plane that divides the data into two classes; there is training speed, high accuracy and a large number of possible applications;
- Decision Tree is the algorithm divides the dataset into small data subsets and builds an associative decision tree for each of them; used to build a model for predicting target values, where prediction rules are built on the basis of previous data; is simple and easy to understand and implement with the ability to explain complex models with clear visualizations, however, it is easily susceptible to overfitting, it performs poorly with non-numeric values, and also shows poor results with small amounts of data.

It was decided to use the Naive Bayes algorithm because it performs well on small amounts of data, is easy to train and operate, and works well with text data. Naive Bayes classifier is a very good option for our system and considering that the number of responses in the dataset is smaller compared to the averages.

4. Results and experiments

To develop this platform, it was decided to use the Python programming language [35-36] and its libraries and frameworks Flask [37], FastAPI [38] and NLTK [39], and javascript and its React library are used for the interface. In order for the user to see data changes on the screen in real time, instead of waiting for complete processing, the Kafka message broker [40] will be used. To create the qualifying part of the feedback classification system, we is proposed to use the Python programming language and the Jupyter Notebook programming environment. To implement the algorithm, it was choosen to use the sklearn library, namely sklearn.naive_bayes.GausianNB. The project also uses the following Python libraries [35-36]: Numpy (working with models), Pandas (data storage and transformation), Re (ribbon manipulation) and NLTK (with tokenize, the TreebankWordTokenizer function for tokenizing words in sentences) and Sklearn (machine learning).

Description of the expanded precedent scenario according to the RUP standard [41-49]:

1. Stakeholders of the precedent and their requirements: The manager wants to receive feedback about the activities of his enterprise or development opportunities;
2. Product user: a manager who will choose the methods and data sources he needs for analysis.
3. Preconditions of the precedent (preconditions):

- The product under development and the payment system must function correctly;
- Developers should find an opportunity to receive data from social networks and Google, as well as process user data;
- The data must be correct;

4. The main successful scenario of the Manager: enter the system → register/authorize → pay the subscription (if it is the first time) → select the required methods → select the required data → receive the results → save the data;

5. Expansion of the main script or alternative streams:

- The manager cannot log in: The system informs the client about an error → The system returns the client to the beginning.
- The manager has entered incorrect data: The manager receives an error message that the data was entered incorrectly → The data is sent to the system again.
- The manager chooses methods: Search by keywords; Sentiment analysis; Popularity of requests; Generalization of the text; Search for optimal locations.
- The manager chooses data sources: Google; Reddit; Twitter; Facebook;
- The manager wants advanced results, so he chooses Graphing and Reporting Development.

6. Postconditions: The manager received the results; Data saved; The manager saved the data;
7. Special system requirements are to ensure the reliability of data transmission, the protection and security of personal data, a convenient interface, round-the-clock support and fast processing of the request.
8. List of necessary technologies and additional devices: The developed product must be a web platform; A device for visual display of results.

On the diagram in Fig. 3 the following issues are shown:

- The manager is the main actor who interacts with the platform; The manager must log in; If it is not possible to enter, the user will receive an error; In order to authorize, the user must register and pay the subscription fee;
- The manager must choose the methods he wants to use: Sentiment analysis, Popularity of requests, Generalization of text, Search for optimal locations;
- The manager must select data sources; Manager can choose to download data from: Google, Facebook, Reddit, Twitter or download own data; If the user enters his own data, and it is incorrect, he will receive a data entry error;
- The manager receives results; The manager has the ability to build graphs and generate a report; The manager stores the data.

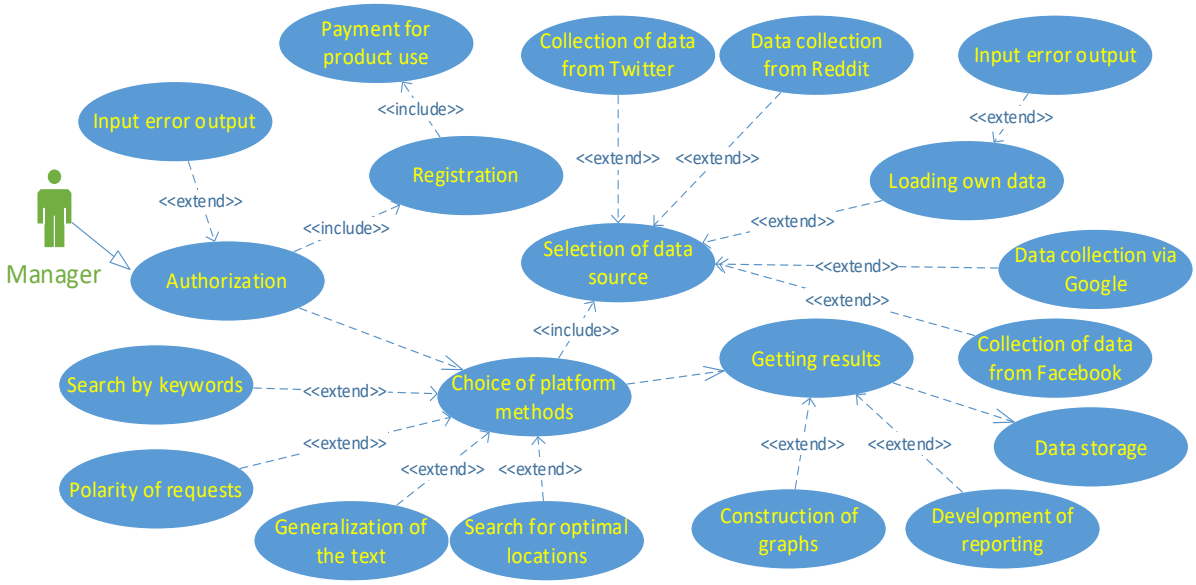


Figure 3: The use case diagram

The diagram in Fig. 4 depicts the main classes and their relationships that make up the successful implementation of the output of our product under development.

1. Registration. This class is used for company registration on the platform, this class is used by the user at the beginning of using the product.
2. Authorization. With this class, a user can log into a company account using a username and password.
3. Data. With the help of this class, the user can choose the source of receiving data.

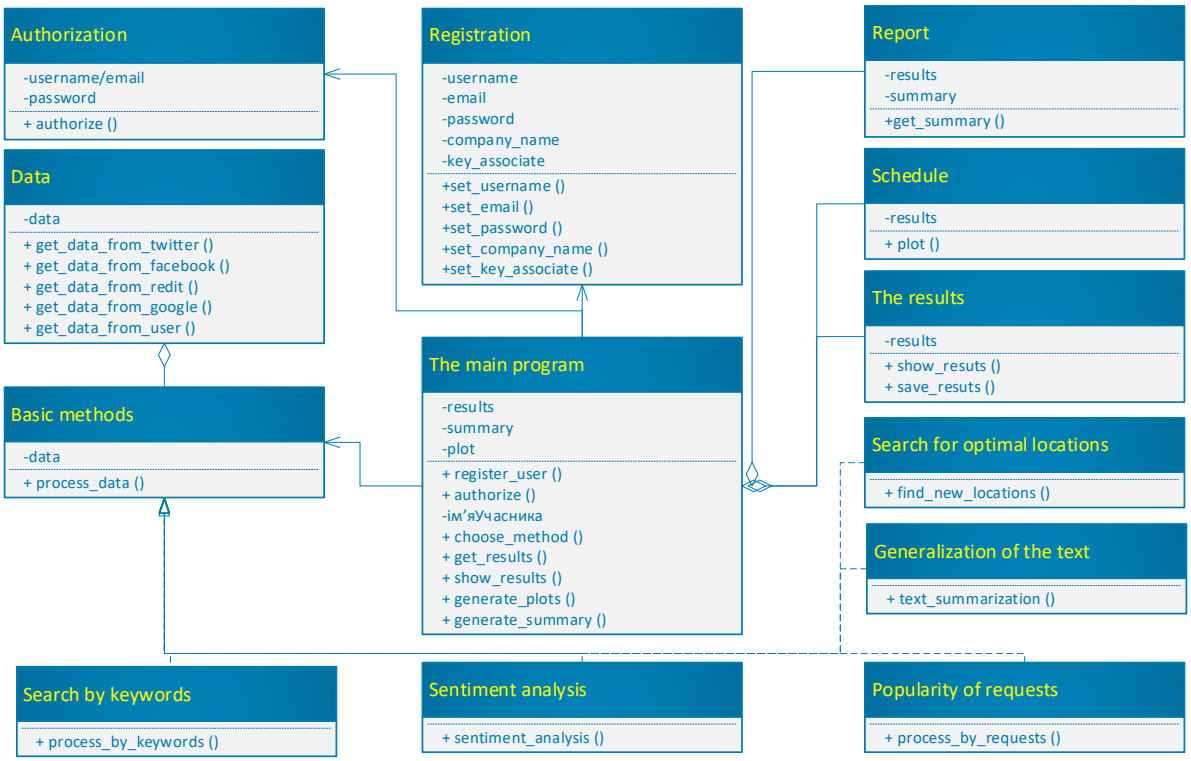


Figure 4: Class diagram

Basic methods (Fig. 5). Keyword Search, Sentiment Analysis, Query Popularity, Text Summarization, Search for Optimal Locations follow the Basic Methods class and contain the implementation of the method of the same name. With the help of the Basic methods class, the object of the Data class is processed, with the help of which the analysis is carried out using NLP methods.

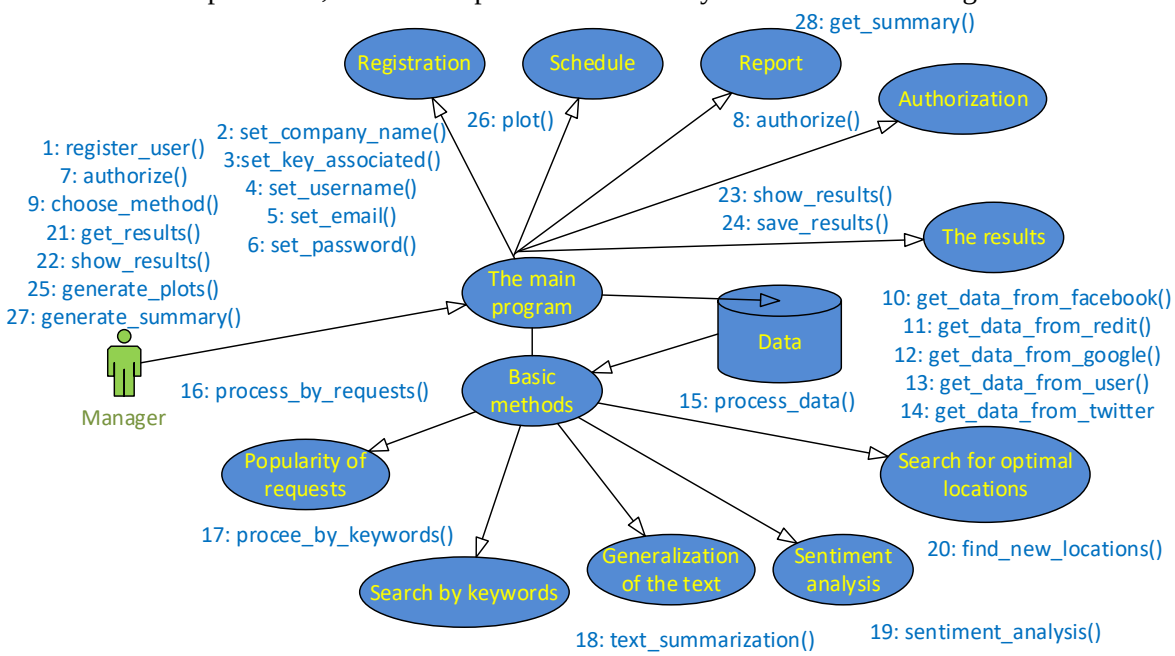


Figure 5: Cooperation diagram

1. Results. With the help of this class, the results are formed, which are shown to the user in the boundary class User screen, and the user can also save the results.
2. Report. With the help of this class, it is possible to generate a report of the results in which the comparison with the previous results is made.

3. Schedule. With the help of this class, it is possible to construct graphs showing the improvement and degradation of customer feedback based on results.
4. The main program. With the help of this class, the system is managed, this class unites all other classes into one system, it forms the workflow of the product.

The diagram in Fig. 6 shows class objects and relationships between them. This is a need to describe the sequence of actions (Fig. 7): to register in the product or to be authorized in the system → to choose a method (Popularity of requests; Search by keywords; Generalization of text; Sentiment analysis; Search for optimal locations;) and data source (Twitter, Facebook, Reddit, Google, own data) → to process data → to get results → to get results on the screen → to save data → to build graphs → to generate a report.

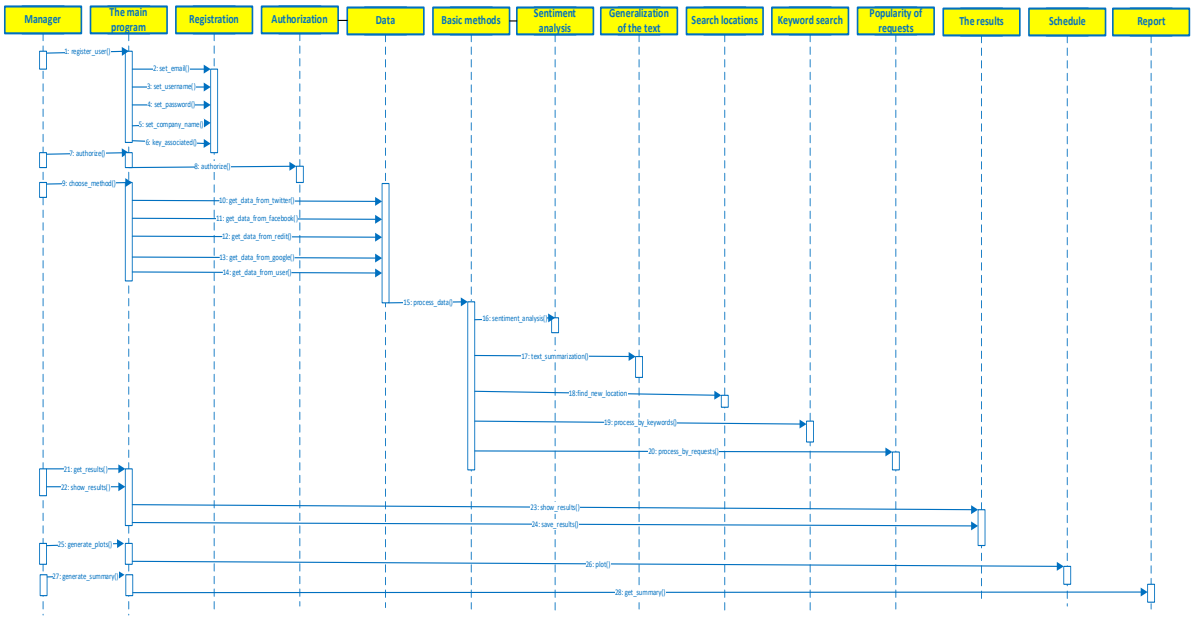


Figure 6: Sequence diagram

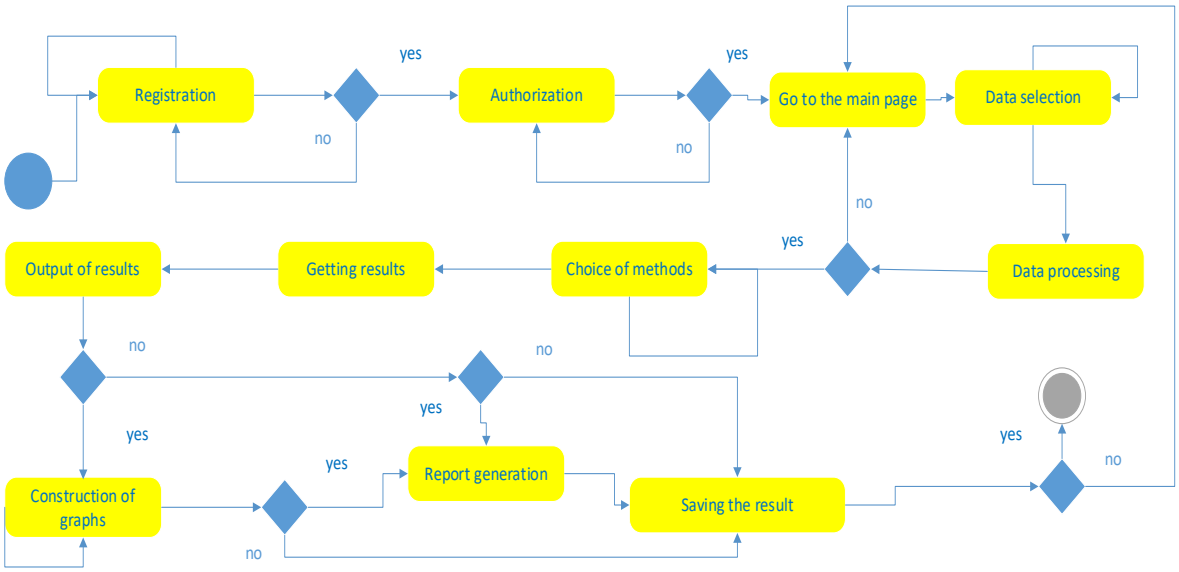


Figure 7: Activity diagram

The diagram in Fig. 8 shows the components that make up the developed program. The interaction between the program components is the following:

- Main.py – this component acts as a leader among components.

- Authorization.py – a component that performs the role of authorization, this component is divided into SignIn – logging into the system, and SignUp – registering into the system;
- DataGathering.py – a component that performs the role of data collection and processing, this component is divided into:
    - Get_data_from_twitter – receiving data from Twitter;
    - Get_data_from_Reddit – receiving data from Reddit;
    - Get_data_from_Google - receiving data from Google;
    - Get_data_from_Facebook – receiving data from Facebook;
    - Get_data_from_user – receiving data from the user;
- Methods.py – a component that contains various NLP methods and others for data analysis. This component is divided into:
    - SentimentAnalysis – sentiment analysis;
    - Search_by_keywords – search by keyword;
    - Popularity_of_requests – popularity of the request;
    - Text_summarization – text summarization;
    - Look_for_new_locations – search for new locations
- Results.py – a component that is responsible for generating the results obtained during the analysis, this component is divided into:
    - Base_results – forms basic analysis results; o
    - Build_graph – construction of graphs based on results;
    - Generate_summary – generation of report results.
    - Save_results – save the results.

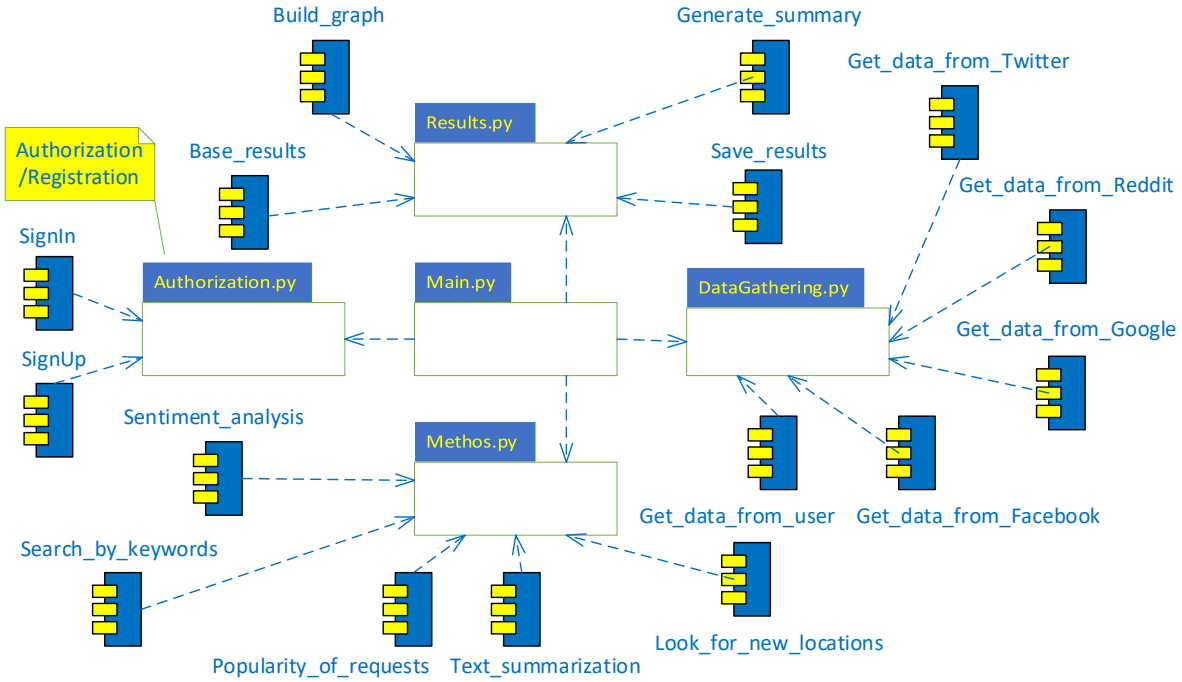


Figure 8: Components diagram

The data is obtained in real time, so it is almost impossible to guess with preprocessing, but it is possible to improve the data received from users, for example, from the social network Twitter. Fig. 9 presents so-called raw data from Twitter with a large amount of garbage not needed for research (for example, many Unicode characters). To do this, based on the re package, it is needed to clear the Unicode data (Fig. 9). A regular expression is a sequence of characters to define a search pattern in text, for example, for "find" or "find and replace" type operations over strings or to validate input data [50]. When clearing the data for each received post, replace all Unicode characters using RegEx and the pattern "[^\x00-\x7F]+" (Fig. 9).

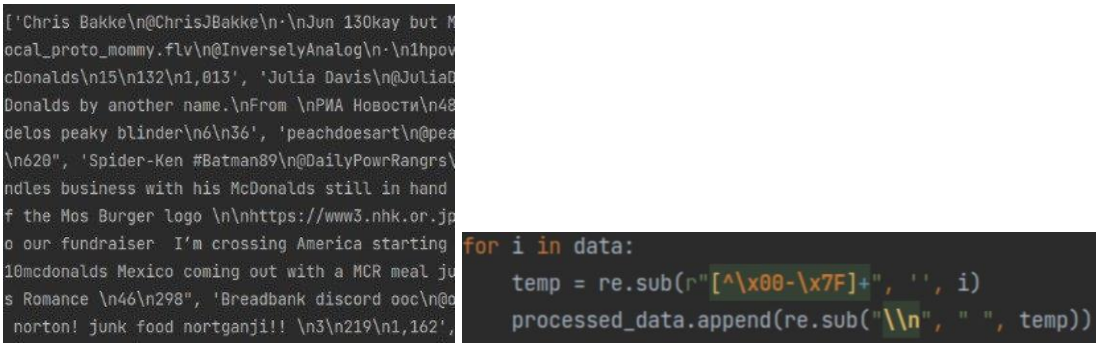


Figure 9: Data sourced from Twitter and data cleaning process

For improved work with this text, it is proposed to tokenize posts using wordtokenization or regexptokenization (regexptokenization works better because it removes unnecessary punctuation marks). Fig. 10 shows the process of tokenization using lemmatization and the data after tokenization.

Figure 10: Lemmatization-based tokenization process and result

NLP methods are developed using Python and corresponding libraries (Fig. 11): Nltk (dataset loading and import of Tokenizer classes), Re (regexp), SentimentIntensityAnalyzer (sentiment analysis), WordNetLemmatizer (lemmatization of sentences into words), PorterStemmer (stemming), Stopwords (dictionary of stop words), Heapq.nlargest (defines the list of n largest elements in the dataset). Fig. 11b shows the loading of additional data for the implementation of NLP methods as a dataset with stop words, punctuation and a Perceptron model for word tagging.

Figure 11: Import of libraries and datasets

In sentiment analysis, the user enters a key according to which he wants to get an estimate of sentiment, for example, the name of the company, product, product category, etc. Next, the data is downloaded from Twitter and processed using regexp. After that, the SentimentIntensityAnalyzer object is initialized, as well as variables for determining the number of positive, negative and neutral posts (Fig. 12). For each post, it is proposed to determine the mood rating and with the help of compound determine to which group the post belongs. If $\leq$-0.05, then negative, and $\geq$ 0.05 – positive, otherwise the post can be considered neutral. After that, to form a percentage distribution and send the data to the client.

```python
def sentiment(topic):
    data = get_twitter_data(topic)
    sia = SentimentIntensityAnalyzer()
    pos = 0
    neg = 0
    neu = 0
    for i in data:
        temp = sia.polarity_scores(i)

        if temp["compound"] <= -0.05:
            neg += 1
        elif temp["compound"] >= 0.05:
            pos += 1
        else:
            neu += 1
    positive = round(pos / (neg + pos + neu) * 100, 1)
    negative = round(neg / (neg + pos + neu) * 100, 1)
    neutral = round(neu / (neg + pos + neu) * 100, 1)
    return {"data": data, "positive": positive, "negative": negative, "neutral": neutral}
```

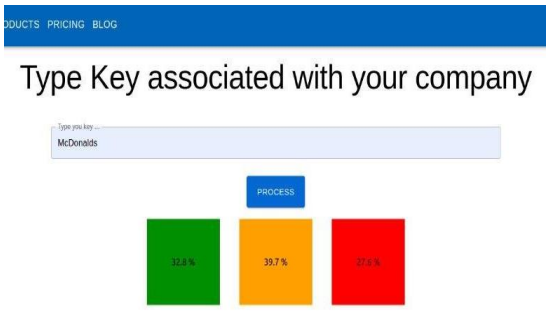
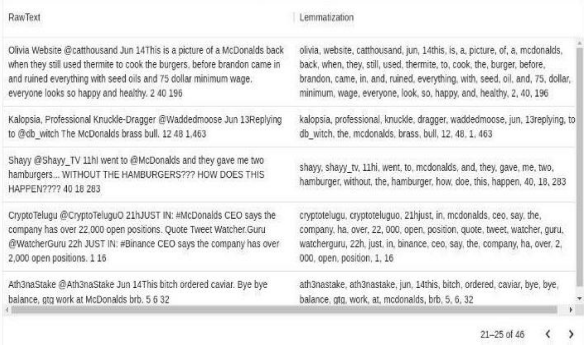


Figure 12: Implementation and result of the Sentiment analysis method

During lemmatization, RegexpTokenizer and WordNetLemmatizer are initialized (Fig. 13). Tokenization is implemented for each post, and a lemmatizer is applied to each token, after which a set of lemmatization results is formed.

```python
def lemmatization(topic):
    data = get_twitter_data(topic)
    wordnet_lemmatizer = WordNetLemmatizer()
    result = []
    tokenizer = nltk.RegexpTokenizer(r'\w+')
    for l,i in enumerate(data):
        tokenization = tokenizer.tokenize(i)
        print(tokenization)
        n = i.strip()
        temp = []
        for w in tokenization:
            temp.append(wordnet_lemmatizer.lemmatize(w.lower()))
        result.append({"id":l,"raw_text": n, "lemmatization": ", ".join(temp)})
    return {"data": result}
```



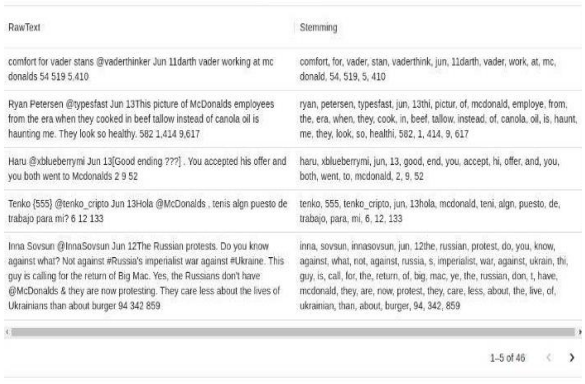
RawText	Lemmatization
Olivia Website @catthousand Jun 14This is a picture of a McDonalds back when they still used thermite to cook the burgers. before brandon came in and ruined everything with seed oils and 75 dollar minimum wage. everyone looks so happy and healthy. 2 40 196	olivia, website, catthousand, jun, 14this, is, a, picture, of, a, mcdonalds, back, when, they, still, used, thermite, to, cook, the, burger, before, brandon, came, in, and, ruined, everything, with, seed, oil, and, 75, dollar, minimum, wage, everyone, look, so, happy, and, healthy, 2, 40, 196
Kalopsia, Professional Knuckle-Dragger @Waddedmoose Jun 13Replying to @db_witch The McDonalds brass bull. 12 48 1,463	kalopsia, professional, knuckle, dragger, waddedmoose, jun, 13replying, to, db_witch, the, mcdonalds, brass, bull, 12, 48, 1, 463
Shayy @Shayy_TV 11hI went to @McDonalds and they gave me two hamburgers... WITHOUT THE HAMBURGERS??? HOW DOES THIS HAPPEN???? 40 18 283	shayy, shayy_tv, 11hi, went, to, mcdonalds, and, they, gave, me, two, hamburger, without, the, hamburger, how, doe, this, happen, 40, 18, 283
CryptoTelugu @CryptoTeluguO 21hJUST IN: #McDonalds CEO says the company has over 22,000 open positions. Quote Tweet Watcher.Guru @WatcherGuru 22h JUST IN: #Binance CEO says the company has over 2,000 open positions. 1 16	cryptotelugu, cryptoteluguo, 21hjust, in, mcdonalds, ceo, say, the, company, ha, over, 22, 000, open, position, quote, tweet, watcher, guru, watcherguru, 22h, just, in, binance, ceo, say, the, company, ha, over, 2, 000, open, position, 1, 16
Ath3naStake @Ath3naStake Jun 14This bitch ordered caviar. Bye bye balance, gtg work at McDonalds brb. 5 6 32	ath3nastake, ath3nastake, jun, 14this, bitch, ordered, caviar, bye, bye, balance, gtg, work, at, mcdonalds, brb, 5, 6, 32

21–25 of 46

Figure 13: Implementation and result of the Lematization method

During stemming, PorterStemmer, RegexpTokenizer and the results variable are initialized as an array (Fig. 14).

```python
def stemming(topic):
    data = get_twitter_data(topic)
    porter_stemmer = PorterStemmer()
    tokenizer = nltk.RegexpTokenizer(r'\w+')
    result = []
    for l, i in enumerate(data):
        tokenization = tokenizer.tokenize(i)
        n = i.strip()
        temp = []
        for w in tokenization:
            temp.append(porter_stemmer.stem(w))
        result.append({"id": l, "raw_text": n, "stemming": ", ".join(temp)})
    return {"data": result}
```



RawText	Stemming
comfort for vader stans @vaderthinker Jun 11darth vader working at mc donalds 54 519 5,410	comfort, for, vader, stan, vaderthink, jun, 11darth, vader, work, at, mc, donald, 54, 519, 5, 410
Ryan Petersen @typesfast Jun 13This picture of McDonalds employees from the era when they cooked in beef tallow instead of canola oil is haunting me. They look so healthy. 582 1,414 9,617	ryan, petersen, typesfast, jun, 13thi, pictur, of, mcdonald, employe, from, the, era, when, they, cook, in, beef, tallow, instead, of, canola, oil, is, haunt, me, they, look, so, healthi, 582, 1, 414, 9, 617
Haru @xblueberrymi Jun 13[Good ending ???] . You accepted his offer and you both went to Mcdonalds 2 9 52	haru, xblueberrymi, jun, 13, good, end, you, accept, hi, offer, and, you, both, went, to, mcdonald, 2, 9, 52
Tenko {555} @tenko_cripto Jun 13Hola @McDonalds , tenis algn puesto de trabajo para mi? 6 12 133	tenko, 555, tenko_cripto, jun, 13hola, mcdonald, teni, algn, puesto, de, trabajo, para, mi, 6, 12, 133
Inna Sovsun @InnaSovsun Jun 12The Russian protests. Do you know against what? Not against #Russia's imperialist war against #Ukraine. This guy is calling for the return of Big Mac. Yes, the Russians don't have @McDonalds & they are now protesting. They care less about the lives of Ukrainians than about burger 94 342 859	inna, sovsun, innasovsun, jun, 12the, russian, protest, do, you, know, against, what, not, against, russia, s, imperialist, war, against, ukrain, thi, guy, is, call, for, the, return, of, big, mac, ye, the, russian, don, t, have, mcdonald, they, are, now, protest, they, care, less, about, the, live, of, ukrainian, than, about, burger, 94, 342, 859

1–5 of 46

Figure 14: Implementation and result of the Stemming method

When summarizing the text of the posts, they are combined into a single array, the stop words of the dataset are marked, initialized and RegexpTokenizer is applied (Fig. 15). For all tokens that do not belong to stop words, it is proposed to create a frequency dictionary, and then normalize the frequency based on the highest frequency found. For each sentence, it is needed to collect the frequency of occurrence of words in other sentences, after which to form a generalization using the nlargest algorithm and combine them into a single whole.

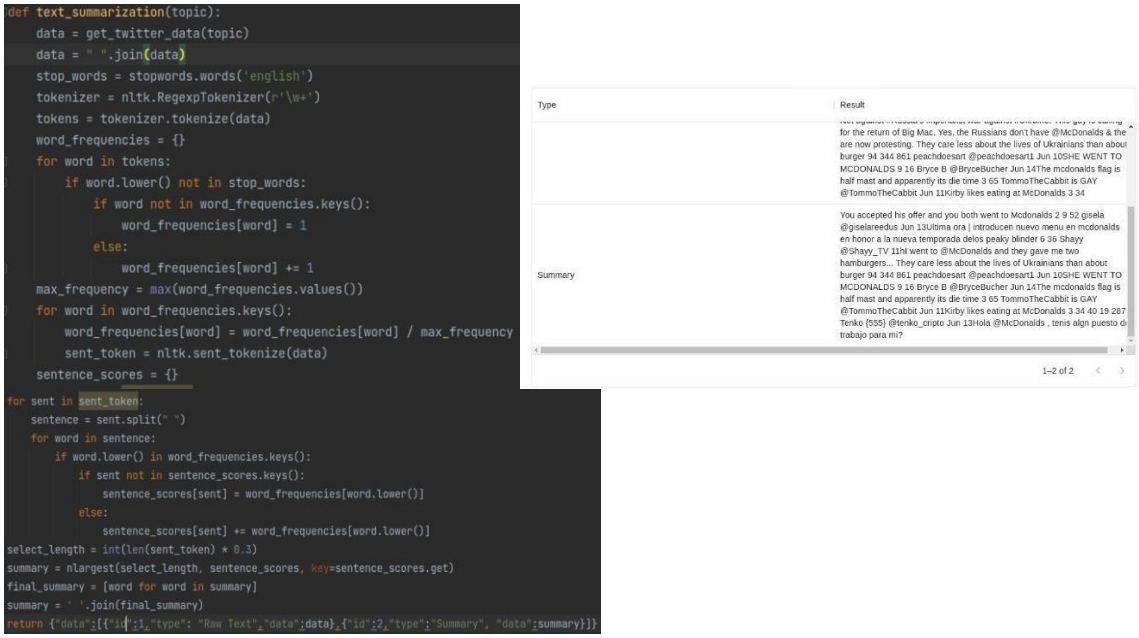


Figure 15: Implementation and result of the text summarization method

When Pos Tagging words, initialize the RegexpTokenizer and the result variable as an array, apply the nltk.pos_tag algorithm (Fig. 16).

Figure 16: Implementation and result of Pos Tagging of words and Tokenization

During tokenization, there is the initialization of the RegexpTokenizer and the result variable as an array (Fig. 17)

Figure 17: Implementation and result of tokenization

There is demonstration of the system with Ukrainian-language texts using the Pandas library (Fig. 18). The feedback dataset is recorded in the form of tsv-falu (punctuation available).

Figure 18: Loading data for training models

After loading the test ones, there is also a need to create an array of stop words (they do not carry any meaning or are excessive noise). The most common stop words in Ukrainian-language posts are "I", "you", "there", "where", etc. (Fig. 19). Also, 'no' is a stop word, but let's exclude this word from the array, since it has a strong enough influence on the value of the response. As part of the classification, the system determines the most important word in the review based on how often that word appears in the Ukrainian language.

```
In [4]: stop_words = ['я',
 'мій',
 'та',
 'сам',
 'ми',
 'наш',
 'самі',
 'ти',
 'ви']
```

```
In [187]: def most_import(review):
              d = {'words': review, 'freq': [0]*len(review)}
              df = pd.DataFrame(data = d)
              for word in review:

                  if word not in set(stop_words) and word != 'не':
                      for i in range(len(word)):

                          for each in freq['word']:
                              if each.startswith(word[:i+1]) and len(each) <= len(word):
                                  refr = each

                          df.loc[df['words'] == refr, 'freq'] = freq['freq'][freq['word'] == refr].iloc[0]

              mst_word = df[df['freq'] == df['freq'].max()]['words'].iloc[0]

              return mst_word
```

Figure 19: An array of stop words of the Ukrainian language and the definition of the most important word in the review

The function checks the first letters of the words and finds the closest match, iterating as many times as there are letters in the word. With each iteration, the number of first letters increases, and at the end, the word from the dictionary that had the largest number of matches is recorded. Further, for optimal classification of reviews, it is necessary to prepare them before training the model based on them. To do this, it is needed to perform a number of operations: remove punctuation; transfer all letters to lower case; tokenization; stemming using the imported Re library, which is designed for working with regular expressions, and using the lower() function, respectively. It is proposed to perform tokenization using the TreebankWordTokenizer function. Stemming is the reduction of words to the smallest possible form, while the meaning of the word is preserved. Stemming is key in any NLP algorithms, as well-executed word reduction allows to optimize the work of later models. The stemming function for the Ukrainian language Ukr_stem has been developed (Fig. 20).

```
In [141]: # def Ukr_stem(review):
          #     stemmed = []

          #     for word in review:
          #         det = 0
          #         bord = 0
          #         word_len = len(word)
          #         found = False

          #         if word not in set(stop_words):

          #             if word_len <= 3 and not found:
          #                 found = True
          #                 stemmed.append(word)
          #             elif word_len == 4:
          #                 for each in slova['word']:
          #                     if each == word:
          #                         found = True
          #                 if found:
          #                     stemmed.append(word)
          #                 else:
          #                     stemmed.append(word[:-1])

          #             else:

          #                 root = word
          #                 for i in range(len(word)):
          #                     for each in slova['word']:
          #                         if i != 0:
          #                             if each.startswith(word[:-i]) and i < (len(word)-3):
          #                                 if len(word[:-i]) < len(root):
          #                                     root = word[:-i]

          #                 stemmed.append(root)

          #     return stemmed
```

```
In [317]: def ukr_stem2(review):
              stemmed = []
              review = [word for word in review if word not in stop_words]
              for word in review:
                  root_len = len(word)-1 if word[-1:] in set(let_1) and len(word) > 2 else 0
                  root_len = len(word)-2 if word[-2:] in set(let_2) and len(word) > 3 else root_len
                  root_len = len(word)-3 if word[-3:] in set(let_3) and len(word) > 4 else root_len
                  root_len = len(word)-4 if word[-4:] in set(let_4) and len(word) > 5 else root_len
                  if root_len == 0:
                      root = word
                      for i in range(len(word)):
                          for each in slova['word']:
                              if i != 0:
                                  if each.startswith(word[:-i]) and i < (len(word)-3):
                                      if len(word[:-i]) < len(root):
                                          root = word[:-i]

                      stemmed.append(root)
                  else:
                      root = word[:root_len]
                      stemmed.append(root)
              return stemmed
```

Figure 20: Functions of stemming Ukr_stem and stemming ukr_stem2

The developed function ukr_stem2 is the second iteration of the stemming function of Ukrainian words. The Ukr_stem function is slow and non-rotating, but accurate. The main idea of Ukr_stem was to compare each word of the feedback with words from the dictionary, which took a lot of time.

The ukr_stem2 function is the best option for obtaining quickly operational data. It checks the endings of words and selects the best abbreviation for it. Arrays containing the most popular word endings in the Ukrainian language have also been created for this purpose. The tree of endings of all possible Ukrainian words developed on the basis of GNU Aspell by Mykola Senyk (Fig. 21) [51] was taken as a sample.

```
In [4]: let_1 = ['я', 'ь', 'и', 'м', 'о', 'у', 'і', 'ю', 'й', 'а', 'є', 'х', 'ї', 'в', 'ш', 'є', 'к', 'т']
        let_2 = ['ся', 'ня', 'сь', 'ть', 'ми', 'ти', 'ли', 'ім', 'им', 'ам', 'ом', 'ям', 'му', 'ну', 'ку', 'мо', 'го', 'ло', 'ні', 'ві',
                 'ті', 'ою', 'ню', 'ій', 'ий', 'на', 'ла', 'ка', 'те', 'не', 'их', 'ах', 'ях', 'ої', 'ів', 'ав', 'еш', 'єш']
        let_3 = ['ося', 'ься', 'ися', 'еся', 'шся', 'ася', 'вся', 'юся', 'ння',
                 'ось', 'ись', 'есь', 'ась', 'всь', 'юсь', 'ють', 'сть', 'ими', 'ами', 'ями', 'ати', 'али', 'нім', 'ним', 'ням', 'ому',
                 'ймо', 'ємо', 'ого', 'ало', 'нні', 'ові', 'сті', 'ною', 'кою', 'сню', 'нню', 'ній', 'ний', 'ала', 'йте', 'єте', 'них',
        let_4 = ['мося', 'лося', 'ться', 'тися', 'леся', 'лася', 'мось', 'лось', 'тись', 'лись', 'тесь', 'лась', 'ість', 'ними',
                 'нням', 'ному', 'ного', 'ості', 'існю']
```

Figure 21: Arrays of Ukrainian word endings

As it is seen from the figures, the new function is much shorter in terms of code and also has much fewer comparisons, what reduces the time to process feedback, since the comparison itself is the longest operation in terms of time. The process of preliminary processing of reviews is given by the prepro function (Fig. 22).

```
In [319]: def prepro(data):
              tokenizer = TreebankWordTokenizer()

              corpus = []
              most_import_word = []

              for rev in data:
                  review = re.sub("[^a-яА-Я-іЇїЇєЄ']", ' ', rev)
                  review = review.lower()
                  review = tokenizer.tokenize(review)
                  most_import_word.append(most_import(review))
                  review = ukr_stem2(review)
                  review = ' '.join(review)
                  corpus.append(review)
              df = pd.DataFrame(data = corpus)
              df['import_word'] = most_import_word
              return df
```

```
In [320]: proc_reviews = prepro(data['review'])
          proc_reviews.columns = ['review', 'import_word']
```

```
In [321]: cv = CountVectorizer(max_features = 250)
          x = cv.fit_transform(proc_reviews['review']).toarray()
          y1 = data.iloc[:, 1].values
          y2 = data.iloc[:, 2].values
```

Figure 22: Function prepro and creation of Bag of Words

The results of preliminary processing of the texts feedback can be used to train the models and create the Bag of Words [52] model, on the basis of which the models will be trained. In this model, a text (such as a sentence or a document) is represented as a bundle (multiset) of its words, neglecting grammar and even word order, but preserving multiplicity. This model is best suited for naive Bayes classifiers. The dataset is divided into training and test sets (Fig. 23) but both models are used. The test set will be only 2% of the entire dataset to maximize model training accuracy.

```
In [322]: x_train, x_test, y_train1, y_test1 = train_test_split(x, y1, test_size = 0.02, random_state = 15)
          x_train, x_test, y_train2, y_test2 = train_test_split(x, y2, test_size = 0.02, random_state = 15)
```

```
In [323]: model1 = GaussianNB()
          model1.fit(x_train, y_train1)

          model2 = GaussianNB()
          model2.fit(x_train, y_train2)

Out[323]: GaussianNB()
```

Figure 23: Dataset distribution and model training

The accuracy of the work of the trained models based on markers for determining the sentiment (negative/positive) of the response (Fig. 24-25) has been investigtated. The overall accuracy of the sentiment error model for the test set is quite satisfactory (92.3%). The trained model classifies positive reviews well, but has some problems with negative ones (Fig. 24). Problems can be due to the fact that

people, especially Ukrainians, do not convey the negativity in the feedback directly, more often with neutral expressions or sarcasm.

```
In [13]: predict1 = model1.predict(x_test)
         print('Predicted Actual')
         print(np.concatenate((predict1.reshape(len(predict1), 1), y_test1.reshape(len(y_test1), 1)), 1))

         Predicted Actual
         [[1 1]
          [0 0]
          [1 0]
          [0 0]
          [0 0]
          [0 0]
          [0 0]
          [0 0]
          [1 1]
          [1 1]
          [1 1]
          [0 0]
          [1 1]]

In [15]: cm1 = confusion_matrix(y_test1, predict1)
         print(cm1)
         accuracy_score(y_test1, predict1)*100

         [[7 1]
          [0 5]]

Out[15]: 92.3076923076923
```

Figure 24: The performance result of the second model and the overall accuracy and error matrix of the second model

Concerning the second model, the results are not so good when classifying actions (Fig. 25). The overall accuracy of the model is not too high (61.5%). From the error matrix, it follows that the most problematic is with those reviews that do not require action.

```
         [1 1]]

In [14]: predict2 = model2.predict(x_test)
         print('Predicted Actual')
         print(np.concatenate((predict2.reshape(len(predict2), 1), y_test2.reshape(len(y_test2), 1)), 1))

         Predicted Actual
         [[0 0]
          [0 0]
          [1 0]
          [1 0]
          [1 1]
          [1 0]
          [1 1]
          [0 0]
          [0 0]
          [1 0]
          [0 1]
          [1 1]
          [0 0]]

In [16]: cm2 = confusion_matrix(y_test2, predict2)
         print(cm2)
         (accuracy_score(y_test2, predict2))*100

         [[5 4]
          [1 3]]

Out[16]: 61.53846153846154
```

Figure 25: The performance result of the second model and the overall accuracy and error matrix of the second model

There is a need to conduct testing with "live" feedback. Firstky, classify the review taken from the Internet and another one written by the authors into models. Both reviews are positive and do not require any additional actions. Both models have correctly classified reviews (Fig. 26).

```
In [18]: rev = ["Чудовий заклад! Приємний песонал, дуже смачна їжа, коктейлі і кальян) На 14 лютог
                "улюблений ресторан, кращого в світі не знайдеш"]
         prot_test(rev)

         Головне слово відгука: лютого;  Позитивне/негативне: 1; Потрібна допомога: 0
         Головне слово відгука: світі;   Позитивне/негативне: 1; Потрібна допомога: 0
```

Figure 26: The result of testing "live" reviews

Also, the system has identified the most important words of reviews. Unfortunately, the speed of processing reviews leaves much to be desired. It took 37 seconds to classify two reviews, which translates to about 18.5 seconds per review. Of course, the length of the response has a big impact on the classification time, since stemming still takes the most time.

5. Discussion

The analysis was performed based on machine learning methods (Fig. 27): naive Bayesian classifier (prediction accuracy 71.13%), logistic regression (prediction accuracy 75.67%) and support vectors (prediction accuracy 72.78%). In Fig. 28 presented classification report modules for measuring the quality of forecasts according to the classification algorithm (comparison of true and false predictions).

Accordingly, true positive TP, false positive FP, true negative TN and false negative FN prediction indicators are used to predict indicators in the classification report [53-63]. In particular, when the case (event) is at TN - negative and predicted to be negative; TP - positive and predicted positive; FN - positive, but expected to be negative; FP - negative, but expected to be positive.

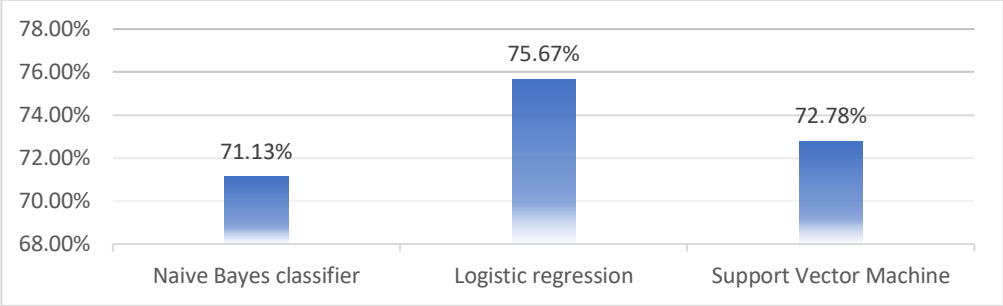


Figure 27: Comparative graph of the results of the used methods

	precision	recall	f1-score	support
negative	0.52	0.54	0.53	128
neutral	0.76	0.87	0.81	575
positive	0.66	0.46	0.54	267
accuracy			0.71	970
macro avg	0.65	0.62	0.63	970
weighted avg	0.70	0.71	0.70	970

71.1340206185567

	precision	recall	f1-score	support
negative	0.70	0.52	0.59	128
neutral	0.78	0.88	0.83	575
positive	0.70	0.60	0.65	267
accuracy			0.76	970
macro avg	0.73	0.67	0.69	970
weighted avg	0.75	0.76	0.75	970

75.6701030927835

	precision	recall	f1-score	support
negative	0.80	0.26	0.39	128
neutral	0.72	0.97	0.82	575
positive	0.77	0.42	0.55	267
accuracy			0.73	970
macro avg	0.76	0.55	0.59	970
weighted avg	0.74	0.73	0.69	970

72.78350515463917

Figure 28: Analysis based on naive Bayesian classifier, logistic regression and support vectors

Precision indicates the ability of the classifier not to mark an instance as positive that is actually negative. For each class, it is defined as the ratio of true positive results to the sum of true and false positive results. This is the accuracy of positive predictions.

Recall indicates the ability of the classifier to find all positive instances. For each class, it is defined as the ratio of true positive results to the sum of true positive and false negative results. So, recall is the proportion of positives that are correctly identified.

F1-score indicates the weighted harmonic mean of precision and recall, where the best score is 1.0 and the worst is 0.0. In general, f1 scores are lower than precision scores because they build precision and recall into the calculation. The weighted mean value of f1 should be used to compare classifier models rather than the global accuracy.

Support indicates the number of cases that correspond to one or another prediction, in our case it is a negative, positive or neutral effect.

6. Conclusions

The application of sentiment analysis of comments, reviews, requests and news for the support and development of e-business has been described. The analyzed analogs made it possible to develop information technology for solving NLP problems of e-business, adapted for the Ukrainian target audience. The general typical structure of the information system for the support and development of e-commerce has been developed by analyzing the feedback of the target audience based on machine learning technology and natural language processing methods. Among the methods of implementing

the main functions, the following machine learning methods are used: naive Bayesian classifier, logistic regression and the method of support vectors. The software was developed and its structure has been described. A review of reports on the implementation of machine learning methods has been carried out. This made it possible to better review and analyze the obtained results. After that, the statistics of the program execution have been carried out, it has been described and the obtained results have been analyzed. Namely, a graph of the comparison of the obtained results was constructed. Also, during the work, a presentation about the developed project has been created and an article has been written, in which the process of working on the project is described in two languages, namely Ukrainian and English. The logistic regression method coped best with the task of analyzing the impact of the news on the financial market, which showed an accuracy of 75.67%. This is certainly not the desired result, but it is the largest indicator of all considered. The support vector method (SVM) coped somewhat worse with the task, which showed an accuracy of 72.78%, which is a slightly worse result than the one obtained thanks to the logistic regression method. And the naïve Bayesian classifier method did the worst with the task, which achieved an accuracy of 71.13%, which is less than the two previous methods. Of course, the obtained results are far from ideal and demonstrate accuracy in the range from 71% to 76%. Which means they need improvement. In the end, I would like to note that this topic is quite popular and relevant, and there are currently no analogues.

7. References

[1] Electronic scientific publication "Effective Economy". Contemporary challenges for the economic development of small business in Ukraine. http://pev.kpu.zp.ua/journals/2021/2_25_ukr/7.pdf

[2] Definition of customer support: https://www.helpscout.com/helpu/definition-of-customer-support

[3] More and more companies outsource parts of their business. URL: http://www.itpaa.org/modules.php?name=News&file=article&sid=2062

[4] Basics of Natural Language Processing for text. URL: https://habr.com/ru/company/Voximplant/blog/446738/

[5] P. Zhezhnych, A. Shilinh, V. Melnyk, Linguistic analysis of user motivations of information content for university entrant's web-forum, International Journal of Computing 18 (2019) 67-74.

[6] A Primer on Neural Network Models for Natural Language Processing. URL: https://jair.org/index.php/jair/article/view/11030/26198

[7] Britannica dictionary. URL: https://www.britannica.com/topic/outsourcing

[8] Sykes. URL: https://www.sykes.com

[9] Sensee. URL: https://www.sensee.co.uk/index.html

[10] Serco. URL: https://www.serco.com

[11] Teleperformance. URL: https://www.teleperformance.com/en-us

[12] Repustate. Using NLP for business success. URL: https://www.repustate.com/blog/using-nlp-for-business-success/

[13] Repustate. How can sentiment analysis help you with Patient Voice? URL: https://www.repustate.com/patient-voice/

[14] SkywellSoftware. How does Siri work: technology and algorithm. URL: https://skywell.software/blog/how-does-siri-workhttps://skywell.software/blog/how-does-siri-work-technology-and-algorithm/technology-and-algorithm/

[15] Grammarly. How Grammarly uses Natural Language Processing and Machine Learning to identify the main points in a message. URL: https://www.grammarly.com/blog/engineering/nlp-mlhttps://www.grammarly.com/blog/engineering/nlp-ml-identify-main-points/identify-main-points/

[16] Klevu. Smart Search Overview. URL: https://www.klevu.com/smart-search/

[17] IBM. Natural Language Processing (NLP). What is natural language processing?. URL: https://www.ibm.com/cloud/learn/natural-language-processing#tochttps://www.ibm.com/cloud/learn/natural-language-processing - toc-what-is-na-jLju4DjEwhat-is-na-jLju4DjE

[18] SaS. (2022). Natural Language Processing (NLP). What it is and why it matters. URL: https://www.sas.com/en_us/insights/analytics/what-is-natural-

languagehttps://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.htmlprocessing-nlp.html
[19] MonkeyLearn. (2020). What Is Natural Language Processing. URL: https://monkeylearn.com/blog/what-is-natural-language-processing/
[20] MonkeyLearn. (2020). Topic Analysis: The Ultimate Guide. URL: https://monkeylearn.com/topic-analysis/
[21] Lexalytics. (2019). Sentiment Analysis Explained. URL: https://www.lexalytics.com/technology/sentiment-analysis/
[22] MonkeyLearn. (2020). Intent Classification: How to Identify What Customers Want. URL: https://monkeylearn.com/blog/intenthttps://monkeylearn.com/blog/intent-classification/classification/
[23] MonkeyLearn. (2020). Keyword Extraction. URL: https://monkeylearn.com/keyword-extraction/
[24] Edia. (2021). What is Keyword Extraction? https://www.edia.nl/keyword-extraction
[25] Towards DataScience. (2022). Stemming vs. Lemmatization in NLP. URL: https://towardsdatascience.com/stemming-vshttps://towardsdatascience.com/stemming-vs-lemmatization-in-nlp-dea008600a0lemmatization-in-nlp-dea008600a0
[26] Analytics steps. (2020). What is Stemming and Lemmatization in NLP? URL: https://www.analyticssteps.com/blogs/what-stemming-andhttps://www.analyticssteps.com/blogs/what-stemming-and-lemmatization-nlplemmatization-nlp
[27] Analytics steps. (2020). What is Tokenization in NLP? URL: https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenizationhttps://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/nlp/
[28] Stanford. (2019). Machine Translation. URL: https://nlp.stanford.edu/projects/mt.shtml
[29] Data Science UA. (2020). Machine Translation. URL: https://data-science-ua.com/wiki/natural-language-processinghttps://data-science-ua.com/wiki/natural-language-processing-nlp/machine-translation/nlp/machine-translation/
[30] Top Coder. (2022). Text Summarization in NLP. URL: https://www.topcoder.com/thrive/articles/text-summarization-in-nlp
[31] Analytics steps. (2022). What Is Text Summarization in NLP? URL: https://www.analyticssteps.com/blogs/what-text-summarization-nlp
[32] Dict_uk Github repository. URL: https://github.com/brown-uk/dict_uk/tree/master/data
[33] Advantages and disadvantages of different classification models. URL: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-classification-models/
[34] Naive Bayes Classifier: https://www.upgrad.com/blog/naive-bayes-classifier/
[35] Coursera. (2022). What Is Python Used For? A Beginner's Guide. URL: https://www.coursera.org/articles/what-is-python-used-forhttps://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-pythona-beginners-guide-to-using-python
[36] Python.org. What is Python? Executive Summary: https://www.python.org/doc/essays/blurb/
[37] PymBook. Introduction to Flask. URL: https://pymbook.readthedocs.io/en/latest/flask.html
[38] FastApi. (2021). FastApi. URL: https://fastapi.tiangolo.com/
[39] NLTK. (2022). Natural Language Toolkit. URL: https://www.nltk.org/
[40] AWS. (2021). What is Apache Kafka? URL: https://aws.amazon.com/ru/msk/what-is-kafka/
[41] Tutorialspoint. (2018). System Analysis and Design – Overview. URL: https://www.tutorialspoint.com/system_analysis_and_design/system_analysis_and_design_overview.htm
[42] Mohammed Maree, Mujahed Eleyat. Semantic graph based term expansion for sentence-level sentiment analysis, International Journal of Computing 19(4) (2020) 647-655.
[43] WeyBackMachine. (2002). System Analysis. URL: https://web.archive.org/web/20070822025602/http://pespmc1.vub.ac.be/ASC/SYSTEM_ANALY.html
[44] Surbhi Bhatia, Manisha Sharma, Komal Kumar Bhatia, Pragyaditya Das, Opinion target extraction with sentiment analysis, International Journal of Computing 17(3) (2018) 136-142.

[45] N. Garanina, E. Sidorova, I. Kononenko, S. Gorlatch, Using multiple semantic measures for coreference resolution in ontology population, International Journal of Computing 16 (2017) 166-176.
[46] Iso.org. (2005). ISO/IEC 19501:2005 - Information technology - Open Distributed Processing - Unified Modeling Language (UML) Version 1.4.2. URL: https://www.iso.org/standard/32620.html
[47] Iso.org. (2012). ISO/IEC 19505-1:2012 - Information technology - Object Management Group Unified Modeling Language (OMG UML) - Part 1: Infrastructure. URL: https://www.iso.org/standard/32624.html
[48] UML.URL: https://web.archive.org/web/20121214050605/http://ooad.asf.ru/Files/U ML.djvu.zip
[49] Tatiana Batura, Aigerim Bakiyeva, Maria Charintseva. A method for automatic text summarization based on rhetorical analysis and topic modeling. International Journal of Computing, vol. 19, issue 1, pp. 118-127, 2020.
[50] Regular expression.URL: https://en.wikipedia.org/wiki/Regular_expression
[51] Tree of endings of the Ukrainian language. URL: http://www.senyk.poltava.ua/projects/ukr_stemming/ukr_endings.html
[52] Bag of Words.URL: https://en.wikipedia.org/wiki/Bag-of-words_model
[53] Understanding the Classification report through sklearn. URL: https://muthu.co/understanding-the-classification-report-in-sklearn/
[54] N. Kholodna, V. Vysotska, O. Markiv, S. Chyrun, Machine Learning Model for Paraphrases Detection Based on Text Content Pair Binary Classification, CEUR Workshop Proceedings Vol-3312 (2022) 283-306.
[55] N. Kholodna, V. Vysotska, S. Albota, A Machine Learning Model for Automatic Emotion Detection from Speech, CEUR Workshop Proceedings Voi-2917 (2021) 699-713.
[56] V. Lytvynenko, M. Voronenko, O. Kovalchuk, U. Zhunissova, L. Lytvynenko, Bayesian Methods Application for the Differential Diagnosis of the Chronic Obstructive Pulmonary Disease, CEUR Workshop Proceedings Vol-2917 (2021) 851-862.
[57] V. Lytvynenko, N. Savina, M. Pyrtko, M. Voronenko, R. Baranenko, I. Lopushynskyi, Development, validation and testing of the Bayesian network to evaluate the national law enforcement agencies' work, in: Proceedings of. 9nd Int. Conf. on Advanced Computer Information Technologies (ACIT' 2019), pp.252-256.
[58] P. Bidyuk, V. Beglytsia, A. Gozhyj, I. Kalinina, Using the Metropolis-Hastings algorithm in Bayesian data analysis procedures, in: Proceedings of IEEE 14th International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2019, pp. 98–101.
[59] P. Bidyuk, A. Gozhvi, I. Kalinina, Modeling military conflicts using Bayesian networks, in: Proceedings of IEEE 1st International Conference on System Analysis and Intelligent Computing, SAIC, 2018, 8516861.
[60] P. Bidyuk, Y. Matsuki, A. Gozhyj, V. Beglytsia, I. Kalinina, Features of application of monte carlo method with markov chain algorithms in bayesian data analysis, Advances in Intelligent Systems and Computing 1080 (2020) 361-376.
[61] R. Yurynets, Z. Yurynets, D. Dosyn, Y. Kis, Risk Assessment Technology of Crediting with the Use of Logistic Regression Model, CEUR Workshop Proceedings Vol-2362 (2019) 153-162.
[62] I. Gruzdo, I. Kyrychenko, G. Tereshchenko, O. Cherednichenko, Applıcatıon of paragraphs vectors model for semantıc text analysis, CEUR Workshop Proceedings 2604 (2020) 283-293.
[63] Zurina Saaya, Tham Weng Hong, The development of trust matrix for recognizing reliable content in social media, International Journal of Computing 18(1) (2019) 60-66.