Вінницький національний технічний університет
____
(повне найменування вищого навчального закладу)

Факультет комп’ютерних систем та автоматики
____
(повне найменування інституту, назва факультету (відділення))

Кафедра автоматизації та інтелектуальних інформаційних технологій
____
(повна назва кафедри (предметної, циклової комісії))

Пояснювальна записка

до магістерської кваліфікаційної роботи
на тему «Інформаційна система інтелектуального аналізу природномовного контенту соціальних мереж. Частина 2. Автоматизована класифікація учасників соціальних мереж на основі лінгвістичного аналізу мікроблогів»

Виконала: студентка 2 курсу,
групи 3АКІТ-18м
спеціальності
151 – «Автоматизація та комп'ютерно-інтегровані технології»
(шифр і назва напряму підготовки, спеціальності)
Освітньо-професійна програма:
Інформаційні системи і Інтернет речей

Стадній О. Ю.
____
(прізвище та ініціали)
Керівник Бісікало О. В.
(прізвище та ініціали)
Рецензент _
(прізвище та ініціали)

Вінниця ВНТУ – 2019 року

Vinnytsia National Technical University

(Full name of university)

Faculty of Computer Systems and Automation

(Full name of institution, name of Faculty (Department))

Department of Automation and Intelligent Information Technologies

(Full name of the department (subject, framing Commission))

Thesis

to Master's qualifying paper
on the topic «Information system of intellectual analysis of natural-language content of social networks. Part 2. Automated Classification of Social Media Users on the Basis of Microblogs' Linguistic Analysis»

Performed by: a student of 2nd course of the 3ACIT-18m group
Specialty:
151 - «Automation and Computer Integrated Technologies»
(Code name and direction of training, specialty)
Educational-professional program: Information Systems and Internet of Things
O. Stadnii
(surname and initials)
Supervisor O. Bisikalo
(surname and initials)
Reviewer ____________
(surname and initials)

VinnytsiaVNTU – 2019

АНОТАЦІЯ

В магістерській кваліфікаційній роботі булорозробленопідсистему лінгвістичного аналізу мікроблогів для автоматизованої класифікації користувачів соціальних мереж. Розробка є складовою частиною загальної інформаційної системи інтелектуального аналізу природно-мовного контенту соціальних мереж.

Було розглянуто теоретичні засади обробки природної мови, представлено математичні моделі аналізу соціальних мереж, дослідженодоступні інструменти та бібліотеки для лінгвістичного аналізу, обґрунтовано вибір тих з них, що найбільш відповідають меті иа завданням дослідження. В результаті виконання роботи було розроблено чатбот для месенджера Telegram, що дозволяє застосувати асоціативні та емоційніознаки тексту мікроблогів для класифікації користувачів мережі.

ABSTRACT

In the master's qualification work a subsystem of linguistic analysis of microblogs was developed in order to design automated classification of social networks’users. Development is an integral part of the general information system of intellectual analysis of natural-language content of social networks.

The theoretical principles of natural language processing were considered, mathematical models of social network analysis were presented, available tools and libraries for linguistic analysis were investigated, the choice of those most appropriate to the purpose and objectives of the study was substantiated. As a result of the work, a chat bot for the Telegram messenger was developed, allowing the use the associative and emotional features of microblogs’ text to classify network users.

CONTENT

INTRODUCTION

The relevance of research. We live in an age when the amount of information produced by a human is greater than ever and the volume of that data is increasing every day. Though, one can gain significant benefit from this information only when properly processed and analyzed [1].

Now, gigabytes of new data of various kinds are being created every moment around the world: new pictures are being taken, videos are being recorded, hundreds of reviews are being made in online stores, thousands are commenting on Facebook entries, dozens of reviews are being made for movies in online cinemas, stock prices are soaring, then they fall. And much of this raw information is practically useless. To get benefit from it, one has to filter and process them. At a time when technology was not so advanced, these operations had to be done manually. It took hours, days, weeks, sometimes months.

Considering the fact that earlier the information for processing was several times smaller, it is easy to understand that it is simply impossible to process such volumes manually. Therefore, many algorithms have been developed that allow these operations to be done using computer technology. Analysis of such methods in relation to natural language processing will be represented in this paper [2].

Social networks are an essential part of the Internet. They have spread our everyday lives and it is rare to meet a person who does not have an account on Facebook, Instagram, Twitter or the forbidden Vkontakte – and even everyone at the same time. It is through these resources that people quickly receive and disseminate information, communicate. But modern social networks are not just a place to chat with friends online. They are a good playground for marketing, sociological research, dating, interest grouping, information, influence, advocacy, intelligence tool for identifying socially dangerous elements and more. User classification algorithms are already widely used by social networks themselves to support user engagement.

The purpose of the study is to improve the classification of social network participants by a specific attribute for a wide range of people based on creating the chatbot that would analyze users' profiles, news and reactions of people in social networks to the important events.

Tasks solved in the work: determining and description of the key approaches of NLP, reviewing existing libraries, developing methods for automated classification of social network users on the base of content and sentimental analysis, development and testing the chatbot as appropriate software.

The object of study is the process of processing information to people classification in the social networks.

The subject of research– methods of automatic classification of participants of social networks.

Research methods: analysis, analogy, modeling, classification and experiment.

The **scientific novelty** is to automate the construction of methods for classifying social network participants. Unlike existing solutions, the proposed development will work on the basis of associativity (words and concepts will be grouped by semantic similarity), which allows improving the quality of classification of social network participants. The other thing that is unique is that everything is managed via chatbot which makes using the software extremely easy and comfortable.

Practical value. Created models, algorithms and software can be used for corporate, government, or personal purposes, for example, observing a specific sample people in sociological research, finding a target group experience in marketing research, identifying psychologically unstable or potentially dangerous individuals for society, etc.

Approbation. Work and some aspects of it, obtainedgeneralizations and conclusions were published on the XLVII scientific and technicalComputer Systems and Automation Faculty conferences (2018) [1] and on the Ukrainian scientific-practical Internet conference "Youth in Science: Research, Problems, Prospects" (2020).

Publications. The results are reported in the abstracts of the XLVII Scientific and Technical Conference of the Faculty of Computer Systems and Automation (2018) [1]and of the Ukrainian scientific-practical Internet conference "Youth in Science: Research, Problems, Prospects" (2020).

1 RESEARCH ON THE CURRENT STATE OF ISSUE

1.1 Key Approaches of Natural Language Processing

Natural language processing is one of the most important areas of research in the field of artificial intelligence. The origin of this trend is associated with the advent of the first computers and with the idea that it would be good to use machine`s to solve useful problems associated with the natural language that people speak and write.

One of the first was the task of machine translation – automatic translation of text from one language to another using a computer or computer. The second task is the creation of dialogue systems that conduct dialogue with a person in a natural language. In the 50th of XX century, the first such programs and systems appeared. The third task is the creation of question-answer systems that accurately answer the question of a person, which is usually presented in the form of text in a natural language.

Today, the relevance of the direction is primarily associated with the need to process large amounts of audio and text information accumulated by mankind over the past decades. The range of tasks for processing natural language has become much wider:

– The task of speech recognition, translating speech into text, which is more convenient to work with: when we talk about working with natural language, we most often work with texts in natural language.

– The inverse problem of speech synthesis.

– Tasks associated with information retrieval, with the extraction of information from texts, when we identify some types of entities in the text and establish relationships between them.

– The tasks of text classification, when it is necessary to attribute the text to one of the previously known classes.

– The tasks of text clustering are grouping multiple texts into groups of similar texts.

– The tasks of summarizing texts when we need to get a concise presentation of it from a large text.

– Tasks related to the analysis of social networks, extracting the opinions of users of social networks about given objects.

– The tasks of creating question-answer systems, dialogue systems, and machine translation tasks also remain relevant today.

We will review this task more detailed in the next subsection.

From a computer science point of view, text is unstructured information and a sequence of characters. To solve the previously described tasks, this must be taken into account, therefore, two fundamentally different ways are possible. The first is to solve problems by considering the text as a chain of characters, bytes, words, and we do not attract any knowledge of the language, although language is a well-structured system that has its own levels, phonetics, morphology, syntax, semantics. The second way is to solve the indicated tasks of processing a natural language with the help of knowledge about what a language is, how it is structured, while we carry out a linguistic analysis of the text.

Usually, a morphological analysis is carried out: the establishment of certain invariant forms of pseudo-bases or lemmas for words, the establishment of morphological characteristics, such as part of speech, number, gender, case. For different languages, the set of morphological characters may vary. Next, parsing is performed: the allocation of syntactic groups or the establishment of syntactic dependencies between words within the sentence. Further, semantic analysis, discursive analysis, and sometimes subject analysis of the text can be performed. Many tasks of processing a natural language can be solved without involving knowledge of what a language is, but most tasks are solved effectively using knowledge of the language and performing linguistic analysis of the text.

At different times, mathematical, logical, statistical, stochastic approaches were used to process the natural language: finite automata, finite converters, predicate logic, Markov chains, formal grammars and probabilistic approaches, and machine learning has dominated since the 2000s.

Today, a typical linguistic study looks as follows: texts that we plan to analyse are selected, a corpus of texts is created, then an expert linguist and an expert in a subject field write some rules or compose dictionaries for identification in texts of target structures. These rules and dictionaries are used to solve the problem.

Another way: a linguist with an expert in a subject field mark texts, mark target constructions in a text, or assign texts to one or another class. Further, machine learning methods automatically derive rules or models that allow to further solve problems; machine learning methods repeat the manual work of linguists and experts. We always check the quality of the work of our methods: there is some reference marked-up sample of texts on which we evaluate quality metrics, how well the machine repeats the work of a person.

One of the main problems of natural language analysis is ambiguity, which manifests itself at all levels and is usually removed due to the context and taking into account some regularity in the use of constructions in the language. The second problem is that the methods of text analysis are highly dependent on the language, genre, subject area: the analysis of literary text is not the same as the analysis of news or text from social networks, so some configuration is always required.

Today, there are quite a few open libraries for processing natural language, and any student who does not have advanced programming skills can use these libraries and solve not-so-difficult text processing tasks. In addition, there are libraries and platforms for processing natural language from industry giants, for example IBM, Google, Microsoft. Natural language analysis is quite widely used in commercial systems, which indicates the high development of this area of research [4].

One of the trends in natural language processing is the use of machine learning methods to reduce labour costs for marking up texts, machine learning methods without a teacher or with partial involvement of a teacher, methods of active machine learning, learning with transfer or indirect learning. The next trend is the use of distributive language models. Vector representations of words and any other language constructs have shown very good efficiency in solving language processing problems, it is also deep machine learning, neural networks. Many tasks of natural language processing today are solved using vector representations and deep learning of neural networks [5].

Recently, the emphasis has shifted to the analysis of social media messages, texts that are written by ordinary people. In addition, natural language analysis is widely used in various subject areas: medicine, psychology, sociology. If we return to the tasks of machine translation and the creation of dialogue systems, then these tasks cannot be finally solved without modelling such concepts as human consciousness, mentality, a picture of the world, without modelling the cultural and historical features of human consciousness.

1. Classification of the text (text classification, text categorization)

The main goal of the classification of the text is to determine the topic of a piece of data. A system gets text and gives it some predefined tag. It is widely usedin:

– sentimental analysis, where the tag contains information about the emotive tint of the source data. Most often there is used 3-tint gradation: positive, neutral and negative;

– spam identification – two categories: spam or not spam – that is widely used by different email services like Yahoo, Gmail etc.;

– language determination – understanding what is the language of the analysed text [6];

– classification of the genre of the book, text, story and so on;

– style classification – determining the style of the message, letter, article etc. i.e. formal or informal;

– categorization of scientific works by domains.

There are also cases when multiple tags are needed to satisfy the task requirements. For example:

– finding the keywords for the text or determining hashtags for tweets or some posts;

– categorization of news by topic and location;

– patience’s heath records in health-care institutions that are indexed by insurance codes, diseases codes, types of procedures.

There are a few approaches to text categorization:

– manual – categorization that is completely done by a human. Accuracy of this approach varies because it completely depends on the qualification of the person that conducts a classification;

– based on hand-crafted automatic categorization rules – categorization is done automatically on the basis of preliminary manually created rules;

– statistical categorization uses machine learning to create automatic categorization rules using training sets, that contain data labelled by human.

2. Language modelling

Usually, language modelling is a part of the bigger NLP tasks like tasks that use a model as an input.

The task is to predict the end of a phrase or sentence based on the beginning of it or on the previous word. Language modelling is a fundamental task for speech, handwriting and optical recognition is used for statistical machine translation and spelling correction [7].

It is also widely used as a part of various natural language processing libraries and architectures.

Linguistic modelling is used to study statistical relationships between language units in order to generate new units' sequences that will be concordant with the target text.

Linguistic models can be used to generate:

- Text headings;
- Pieces of text like sentences;
- Pieces of speech;
- Abstracts etc.

3. Speech recognition

The task of speech recognition is to extract a sense of utterance. It is done by converting the sound that contains natural language into the words that correspond to what has been said.

So input is audio file and output – text data.

There is a specific field for it which name is ASR – automatic speech recognition [8].

- Examples of speech recognition:
- Voice command of a car or smart home;
- Decoding of the speech;
- Subtitles to movies and shows;

4. Creation of the captions

The captions should describe what is the concept of some image or video.

As an input, it gets some digital image and as output, it gives us a text with the description of its content.

Examples:

- description of photos;
- description of some movie scenes;
- captions of videos.

It helps to improve browsing on the Internet and makes finding these photos or videos easier since created text description can be searched. It also allows people with visual or aural impairments to use the Internet and to watch the movies [9].

5. Machine translation

Machine translation is an automated transformation of the speech or text from one language to another. It is one of the biggest fields of NLP and one of the most important spheres of its usage. It may accept and generate both audio and textual data and transform it from one language to another. Machine translation is probably the oldest NLP task. It involves all the modern technologies, including machine learning [10].

6. Summarization of documents

The aim of summarization is to create a concentrated concept of a source text that will reflect its main idea and will give a clear understanding what this text is about.

In other words, it is a creation of a short resume of a text. There are two types of approaches that are used in modern summarization systems:

- extractive – cutting off and stitching elements of the text together in order to eventually create a consistent extract. It is quite limited method. Most of such approaches rely on the human-engineered traits (i.e. surface traits like sentence position and length, words that are present in the title, occurrence of proper nouns in the text; content features like frequency of word in the text; event features like action nouns etc.), although there also are data-driven ones, some of which are based on neural networks [11];
- abstractive is completely opposite approach that aims to create a summary that may not contain parts of the source text; it is concise

paraphrasingof the document’s content that uses vocabulary, that is not used in the input document. We literally map words’ sequence from input text to resulting sequence that is a summary. Neural networks are also very widely used in abstractive approach models.

The result – summary – is usually really short and it does not much depend on the size of the original text. The main idea is to compress the sense of the source as much as it is possible losing some details to get the key concept of it disregarding minor details – that is what makes the major difference between summarization and machine translation although they have a lot features in common [12].

We can also distinguish different levels of summarization: sentence-level, paragraph-level, text-level etc.

The examples of summarization:

– creation of a document title;

– generating an abstract of a document.

These tasks can be divided into levels (mentioned in brackets). We distinguished:

– Recognition of text, speech, speech synthesis (signal);

– Morphological analysis, canonization (word);

– POS-tagging, recognition of named entities, word highlighting (collocation);

– Parsing, tokenization of sentences (sentence);

– Extracting relationships, defining language, analysis of emotional coloring (paragraph);

– Document abstract, translation, subject analysis (document);

– Deduplication, information retrieval (building) [13].

1.2 What Is Social Network

A network is a set of sites (such as people, organizations, web pages, or government entities). Also, these are relationships (or connections between these nodes). Each relationship connects several nodes. If the relation is directed, it is represented as a vector; if not directed – is represented as an edge. The postal network, for example, is a directed network of senders and recipients. A social network organized using software is usually a non-directed network of friends (users).

A prerequisite for the emergence of network analysis is the fact that networks are structures that can restrict or promote public communications. For example, if there is only one node that connects two groups, then that node is especially important in information transmission: it can even control information as information flows through it. In addition, networks are interesting structures in terms of their properties: an illustration of all mail network connections produced will be a clear and understandable pattern of relationships for the owner of this mailbox.

Contrary to the postmodern understanding of the network entity, expressed in particular by Bruno Latour's Actor-Network Theory or Deleuze-Guattari's "rhizome", network analysis works best when all the nodes in a network belong to one class of objects. For example, if a blog can have more than one author, there will be enough data to analyze the blog network about the blogs themselves, rather than all of their authors or related web pages. To explore more than one object type (such as bloggers and commentators), you can use "two-tier analysis", which involves your own set of differences. The relationships in this case should be of the same type as the nodes. If the researcher links the email addresses, he should not build a network in the case where one relation indicates "is in the user's address book “A” and the other –

"send an email to user “A’”. Such simplifications of social relations to the only types of nodes and relationships allow the researcher to draw complex networks on a holistic picture of relations between people.

Depending on the purpose of the study, it is possible to build either a very large but superficial social network, or a series of small but detailed networks. The following section focuses on three types of networks. It will illustrate the approaches of sociological research, the discovery of sociologists both online and offline in relation to each of the three types of networks.

1.3 Types of Social Networks

There can be distinguished the following types of social networks:

1. Solid networks.

Whole networks are relationships of people within defined boundaries. Examples of such objects used in online research are: emails, mailing lists, social networks of the Internet space (such as MySpace), people working in the same office connected by local networks. Whole networks are the most commonly used objects in social network analysis. However, these objects cannot always be used. They may vary depending on the practical requirements of the researcher; collecting office-sized network information is not a particularly difficult task, but obtaining a valid list of all MySpace users is virtually impossible, as this list will change as you collect data. In the study of whole networks, the researcher asks questions about group structure, specific types of network users, about outstanding users. Electronic methods allow the researcher to collect fairly modest data over entire networks. For example, one might consider the structure of mutual responses of registered participants in a teleconference. The work of Smith and his colleagues in a Microsoft study illustrated that some teleconference groups have particularly distinguished individuals altruistically answering other participants' questions, while other groups have a free discussion structure.

Whole networks can be actively built. This is usually done on a list. In such circumstances, anyone can contact any member of the network and ask him or her for connections with all other members listed. Each list is a row in a matrix (often in a large format table) that can be used to construct arcs from respondents to all other participants. Active data collection can be useful when assessing the subjective status of participants or when it is necessary to clarify their relationship to the network as a whole; whereas hidden data collection is suitable for behavioral lines research.

2. Ego networks

The main purpose of whole network analysis is to identify the features of the network and to identify why some people occupy a special position in it. For example, is there a complex subgroup in this network, or why are some individuals always answered? In contrast, the analysis of ego-networks is comparative in nature. It determines the difference in size, shape and quality of various ego networks. Ego networks are usually represented by a spontaneous sampling of users. Moreover, the researcher always strives for the representativeness (even ephemeral) of such a sample. Each sampling element in this analysis is referred to as "ego" and the nodes associated with ego are referred to as "other". The researcher can collect data from either a star-shaped network (ego-node and its connections with other nodes), or a complete ego-network (which also includes connections of other nodes with each other).

Researcher can discover ego-network in social networks of the Internet. In this case, it will be a list (such as a list of friends). Further, during the analysis of the list it will be possible to understand who is attached to whom.

Data collection on ego-networks can be based on the already available results of various research techniques and interviews. The most famous of these is the name generator and the position generator. Methods such as the resource generator and the summation method are also used. All of these methods, with the exception of the name generator, have not been fully developed and do not allow us to collect data on the links between "others".

3. Incomplete networks

Incomplete networks are, in fact, a sample of relative data created by the snowball method. Incomplete networks are a trade-off between the desire to reach a whole network and the fact that some whole networks are simply too massive to be completely covered. The researcher can start from a single web page or multiple pages (so-called "sowing"), then he will search for pages related to that sowing and then pages related to those pages. The sampling process ends when enough pages are collected; when all possible pages are assembled; or when the sample matches a certain criterion (for example, when all pages with more than 400 words are collected).

Incomplete networks are a practical solution for analyzing a fairly large body of data on the Internet. A researcher cannot collect information about all blogs, or MySpace users, but he can build a network of relationships that connects the ego networks of many users. Since it is easier to use the snowball technique on the Internet than it is offline, we can assume that the number of researchers using incomplete networks to answer questions about online behavior will grow before their eyes. Currently, the use of incomplete networks is an emerging research area, also called citation analysis.

As each researcher works outside the crop, the use of incomplete networks is due to some concern about the generalizability of the study results. As Rothenberg points out, the selection of pages by means of a snowball in social networks “in the absence of probabilistic sampling, in the case of an unsuccessful statistical superstructure, and, in principle, in the absence of desirable statistical properties, is not acceptable for the researcher”. This, on the one hand, limits statistical generalizations, but on the other hand, does not reduce the descriptive analysis and its conclusions. Thus, generalizations are made not at the statistical level but at the theoretical level. In addition, the researcher can collect most of the material through successfully organized sowing and examine all the connections that meet certain conditions (such as the presence of a specific set of keywords).

Currently, in the analysis of social networks,there are four main areas of research: structural, resource, regulatory and dynamic.

1. In the structural approach, all network members are considered as vertices of the graph that affect the configuration of edges and other network members. The focus is ongeometric shape of the network and the intensity of interactions (the weight of the edges), therefore, we study such characteristics as the mutual arrangement of vertices, centrality, transitivity of interactions. Structural and network exchange theories are used to interpret the results in this direction.

2. The resource approach examines participants' opportunities to leverage individual and network resources to achieve specific goals and differentiates participants,who are in identical structural positions of the social network, by their resources. Knowledge, prestige, wealth, race, gender can serve as individual resources. Network resources are influence, status, information, capital.

3. The regulatory direction examines the level of trust between participants, as well as the norms,rules and sanctions that affect the behavior of participants in the social network and processestheir interactions. In this case, the social roles that are associated with a given edge of the network are analyzed, for example, manager-subordinate relationships, friendships or kinship relationships. The combination of individual and network resources of the participant with the rules and regulations in force in this social network, forms his "network capital". In simplified form, "network capital" can be considered as the sum of some benefits that a participant can receive at any time in order to achieve a certain goal.

4. Dynamic approach – a direction in the study of social networks, in which the objects of research are changes in the network structure over time: for what reasons the edges of the network disappear and appear, how the network changes its structure in external influences, whether there are any stationary configurations of the social network and others.

Let us consider a little more in detail the directions of analysis of social networks in terms of the tasks to be solved [14].

1.4 Social Media Analysis Approaches

Structural analysis and analysis of social networking behavior are needed to identify the most important peaks, connections, communities, and developing regions of the network. This analysis allows us to review the global evolutionary behavior of the network. Structural and link behavior analysis uses statistical analysis methods, community definition methods, and classification algorithms.

Statistical analysis of social networks. "Typical" social networks andthe question is what the network will look like if it is enlarged. The mutual behavior of the vertices of the network is studied based on the assumption that most vertices have little connection, whether "nuclei" (clusters) or the degrees of vertices are distributed more evenly. The behavior of vertices in clustering is studied. Another of the issues studied in this section is the behavior of typical temporal characteristics of social networks. For example, how the network structure changes as it grows or how the behavior and distribution of connected graph components change. Over time, new entities are added to the network, but some properties of the graph may still persist [15].

Identifying communities in social networks is the most important in the analysis of social networks, though quite close to the task of classification. The goal is to tryidentify the regions of the network within which the active interaction of participants.

Algorithmically this problem can be attributed to the problem of graph separation. It is necessary to divide the network into dense regions based on the behavior of the connections between the vertices. Computer social networks are dynamic, leading to difficulties in identifying communities. In some cases, it is possible to integrate the information content of the network into the process of identifying communities. Content is then an aid to identifying groups of participants with similar interests.

One of the best-known examples of network analysis was conducted in the 1970s. American sociologist Mark Granovetter. He showed that for many social tasks, such as job search, weak connections are much more effective than strong ones. He called this effect "the strength of weak bonds."

For the effectiveness and sustainability of a social network, the functional roles of its participants, such as information brokers, experts, and others, are important. Information brokers – people who play the role of an intermediary in a social network, connecting groups of people among themselves, establishing links between specialists and thereby opening them access toinformation. To identify experts in social networks, for example, the optimization approach of the ant colony (ACO – Ant Colony Optimization) is used. This is an effective polynomial algorithm for finding approximate solutions of the optimal path in a graph.

The original idea of the algorithm comes from observing ants in the process of finding the shortest path from the colony to the power source. Biologists have noticed that among experiments on choosing between two paths of unequal length leading from a colony to a power source, ants, as a rule, use the shortest route. The model of this behavior is as follows. The ant passes randomly from the colony. If he finds a source of food, he returns to the nest, leaving a trail of pheromones behind him. These pheromones attract other nearby ants, which are likely to follow this route. Returning to the nest, they will strengthen the pheromone path. If there existtwo routes, then more ants will have to go along the shorter route at the same time than the long one. Therefore, a short route will become more attractive, and long routes, in eventually disappear due to the evaporation of pheromones.

Work begins with the placement of ants at the vertices of the graph, then the movement of ants begins. The direction is determined by a probabilistic method based on a formula (1.1):

$$P_i = \frac{l_i^q \cdot f_i^p}{\sum_{k=0}^{N} l_k^q \cdot f_k^p}, \qquad (1.1)$$

whereP$_i$ is the probability of transition along the path i;

l$_i$is the reciprocal of the weight (length) of the i-th transition;

f$_i$is the number of pheromones at the i-th transition;

q is the value that determines the "greed" of the algorithm;

p is the value that determines the "herd" of the algorithm and q + p = 1.

The solution is not accurate and may even be one of the worst, however, due to the likelihood of a solution, repeating the algorithm can produce a (fairly) accurate result [16].

Another widely known example of the analysis of social networks is the experiment of the American psychologist Milgram, conducted in 1969. This experiment was called“The phenomenon of the small world” (Milgram's small world experiment), or the “theory of six handshakes” (Six degrees of separation). The hypothesis is that each person is familiar with any other inhabitant of the planet through a chain of common acquaintances, on average consisting of six people. So far this claim has not been refuted. On the contrary, as evidence of the correctness of the hypothesis, an observation is put forward that the diameter of most networks is relatively small [17].

Among the methods of analyzing social networks, the main ones are: graph theory methods, in particular the study of oriented graphs and their representing matrices, used to study the structural relationships of a network participant; methods for finding the local properties of participants, for example, centrality, influence, position, belonging to certain subgroups; methods for determining the equivalence of participants, including their structural equivalence; block models and role algebras; analysis of dyads and triads; probabilistic models, including models of Markov processes.

Graph models. Any social network can be mathematically represented as a graph (formula 1.2):

$$G = (V, E), \qquad (1.2)$$

where V is the set of vertices of the graph;

E is the set of edges of the graph;

|V| = N is the quantity of vertices in the graph.

In the graph of a social network, the vertices are the participants, and the edges indicate the presence of relations between them. Relations can be either directional or non-directional.

Usually, there are two types of relations: “friendship” (people are familiar with each other) and “interests” (there are common interests, people are in the same group of interests). This relationshipused, for example, in FOAF (Friend of a fried) – an ontology for describing people, their activity and relationships to other people and objects. At FOAF, the description of social relationships between people is based on the transitivity of trust. A description of the algorithm for calculating the level of trust (TrustRank) is given below.

Three types of graph models can be distinguished.

1. Stochastic block models are defined by a matrix A of size N × N, where N is the number of groups (blocks) of participants. The element $a_{ij} \in 0,1$ shows the density of connections between network members belonging to the group $v_i$and participants belonging to the group v_j.Moreover, the graph does not contain additional edges and vertices corresponding to the bonds of the participants within the same group [18].

2. Probabilistic graph models are defined by a matrix A of size N × N, where N is the number of network participants. The element $a_{ij} \in 0,1$ shows the probability of interaction of the participant $v_i$andparticipant $v_j$for a certain period of time.

3. Ordinary graph models are defined by a connection matrix A of size N × N.

For the analysis of graph models of social networks, it is sometimes convenient to use a density coefficient, defined as the ratio of the number of edges in the analyzed graph to the number of edges in a complete graph with the same number of vertices (a complete graph is a graph in which all vertices are connected to each other). In addition, the network can be characterized by such quantities,as the number of paths of a given length (a path is a sequence of vertices interconnected), the minimum number of edges whose removal splits the graph into several parts.

Graph models of social networks are used to model the economic and communication relationships of people, analyze the processes of disseminating information, finding communities and related subgroups into which the entire social network can be divided [19].

Analysis of centrality and other local properties. To determine the relative importance (weight) of the vertices of a graph (i.e., how influential is a participant in a particular network), introduce the concept of centrality – a measure of proximity to the center of the graph. Centrality can be defined in different ways, so there are various measures of centrality. It should be noted that this is not about geometric centrality when visualizing the relationship graph.

Degree centrality is defined as the number of connections incident to a vertex (formula 1.3):

$$C_D(v) = \deg(v) \qquad (1.3)$$

where $C_D(v)$ – degree centrality;

v – vertex.

Allocate incoming and outgoing connections. Incoming communications characterize the popularity of a person; outgoing communications characterize his sociability. The resulting value can be normalized by dividing by the total number of participants in the network.

In other words, centrality suggests that among network members, the one with the most friends, or the one with more communities, is more influential. However, a network member with a large number of friends can be associated with the rest of the graph with a small number of edges. Therefore, the following concept is introduced [20].

Closeness centrality is an indicator of how quickly information spreads across the network from one participant to the rest. As a measure of the distance between two participants used the shortest path along the graph (geodetic distance). So, the participant's immediate friends are at a distance of 1, friends of friends are at a distance of 2, friends of friends of friends are at a distance of 3, etc. Next, the sum of all distances is taken and normalized. The obtained value is called the distance of the vertex v from other vertices. Proximity is defined as the reciprocal of the distance (formula 1.4):

$$C_C(v) = \frac{N-1}{\sum_{t\in V\backslash v} d_G(v,t)}, \tag{1.4}$$

whered$_G$(v,t) is the shortest path from the vertex v to the vertex t.

In other words, centrality in proximity allows you to understand how close the member in question is to all other network members. Thus, it is important not only to have direct friends, but also that these friends themselves also have friends [21].

Centrality through mediation (Betweenness centrality). Another characteristic of a participant is its importance in disseminating information. It is in this context that the centrality of mediation evaluates the participant. It is calculated as the number of shortest paths between all pairs of participants passing through the consideredmember (formula 1.5):

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{1.5}$$

whereσ_{st} is the total number of shortest paths from vertex s to vertex t;

σ_{st}(v)is the number of shortest paths from vertex s to vertex t passing through vertex v.

For normalization, it is necessary to divide by the number of pairs of vertices, with the exception of the vertex v itself, i.e., for a directed graph, divide by (N – 1) (N – 2); for an undirected graph, by a quantity equal to(N – 1) (N – 2) / 2. The disadvantage of centrality in mediation is its computational complexity [22].

Own vector centrality (eigenvector centrality). Let the centrality of the participant under consideration be x_v and the centrality of his immediate friends (neighboringvertices)x_j, x_k, $x_l$and so on. Eigenvector centrality is defined as the sum of the centralities of neighboring vertices divided by the constant λ (formula 1.6). Having written similar equations for all friends, we obtain a vector of unknowns (formula 1.7)

$$x_v = \frac{(x_j + x_k + x_l)}{\lambda}; \tag{1.6}$$

$$X = (x_l, \ldots, x_v, \ldots, x_n). \tag{1.7}$$

Addition rules are determined by the adjacency matrix A = (a_{vt}), meaning a_{vt} = 1, if the vertex v is connected to the vertex t, and a_{vt} = 0 in the other case.

Next, it is needed to solve equation AX = λX, that means to find the eigenvalues and eigenvectors of the matrix A.

The resulting task can be rewritten differently, like in the formula 1.8:

$$C_E(v) = x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{vt} x_t, \tag{1.8}$$

where M(v) is the set of vertices adjacent to the vertex v;

λ is a constant.

The eigenvector corresponding to the largest eigenvalue is just formed by the centralities of the corresponding network participants [23].

Thus, the more friends the participant has and the more central they are, the greater their centrality. The converse is also true: the greater the centrality of the participant, the greater the centrality of his friends. The disadvantage of centrality in its own vector is also computational complexity.

A generalization of centrality in degree is Katz centrality.The difference is that centrality by degree takes into account the number of immediate neighbors of the vertex, and centrality of Katz takes into account the number of all vertices that can be connected by the path (formula 1.9):

$$C_{Katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (a^k)_{ji}, \tag{1.9}$$

where $\alpha \in (0.1)$ is the participation fraction of distant vertices, called the attenuation coefficient.

Katz's centrality can be represented as a kind of centrality in its own vector (formula 1.10):

$$C_{Katz}(i) = x_i = \alpha \sum_{j=1}^{N} a_{ij} \left(x_j + 1\right). \tag{1.10}$$

Centrality can be calculated using a link ranking algorithm(PageRank), which is used by the Google search engine. It is based on the principle of "importance" of a web page: the more links to a page, the more important it is. Besides,the weight of the page itself is determined by the weight of the link sent to it by the page. Thus, PageRank is a method of calculating page weight by counting the importance of links toit, i.e., a vertex that refers to another vertex with a large weight, itself receives a large weight (formula 1.11):

$$C_{PageRank}(i) = x_i = \alpha \sum_{j=1}^{N} a_{ji} \frac{x_j}{L(j)} + \frac{1-\alpha}{N}, \tag{1.11}$$

where L(j) = $\sum_j a_{ji}$ – the number of vertices adjacent to the vertex j (or the number of outgoing links in the directed graph).

The difference between this algorithm and the calculation of centrality by the eigenvector and centrality of Katz is in the presence of a conversion factor

L(j). It should also be noted that the reference ranking algorithm uses the inverse indexing of the adjacency matrix a_{ji}in comparison with the calculation of centrality by the eigenvector. The predecessor of the PageRank algorithm is the HITS (Hyperlink-Induced Topic Search) algorithm proposed by Kleinberg.

In addition to the listed methods for determining centrality, there are a large number of methods for calculating this network characteristic introduced in a non-classical way [24].

Important characteristics of network connections are balance and transitivity.

Balance is the absence of situations like “positive interaction (friendship, partnership) between v_1 and v_2, as well as between v_1 and v_3, but negative interaction (hostility, rivalry) between v_2 and v_3”. It is argued that balanced networks are psychologically more comfortable for participants and more stable compared to unbalanced ones. Transitivity is the fulfillment of conditions of the form “if there is an interaction between v_1 and v_2, as well as between v_2 and v_3, then there is an interaction between v_1 and v_3”. These characteristics describe the local connections of participants and are often used in the analysis of dyads and triads.

A useful feature when analyzing social networks is the level of trust. An algorithm for calculating the level of trust (TrustRank) was proposed M. Davern. It was originally created for separation of informative web pages from spam. If we talk about this algorithm in terms of sites, for a control sample, experts manually evaluate the degree of confidence of a small number of sites that can be considered reliable. These sites are taken as a reference. Further, the algorithm is based on the assertion that good sites rarely refer to bad ones, but bad ones very often refer to good ones. TrustRank– a value that gives an assessment of whether it is possible to trust a particular site, considering that it does not containspam. The more links on the site, the less trust is “transmitted” to each such link. The degree of trust to the site (TrustRank) decreases with increasing distance between it andinitial sampling.

The strength of the participant’s structural position is the main indicator determining the differences in the resources of network participants. In the theory of network exchange, to measure this characteristic, the participant's GPI strength index v_i (Genuine Progress Indicator) is introduced in formula 1.12:

$$GPI_i = \sum_{k=1}^{g-1}(-1)^{k-1}P[1]_k, \qquad (1.12)$$

where$P[1]_k$ is the number of disjoint paths of length k passing through the vertex v_i.

Force of the participant $v_i$compared to the strength of participant v_j is calculated as in formula 1.13.

$$GPI_{ij} = GPI_i - GPI_j. \qquad (1.13)$$

2 TOOLS USED IN APPLICATION DEVELOPMENT

2.1 Libraries for NLP Application Development

Natural language processing is a technology thatis used in voice assistants, chatbots, in all applications that process and predict text, speech – and it literally seeped into almost every sphere of our lives, especially in last few years, when it expanded tremendously. There is a great choiceof open-source tools that can be used to solve various NLP tasks. Every popular programming language has its own ones. We will review the most widely used languages and NLP libraries that are available for themin order to choose the best tool for current research.

Natural language processing is the confluence of Artificial Intelligence (AI) and data science. The basic principle is to make machines to get the meaning of what is being said or written. For this purpose, NLP projects often are using a machine learning.

The reason why NLP is so popular nowadays is the possibility to gather valuable customers' data that will be analysed to, for example, create better marketing strategy or to help in choosing the best option for some product. It helps to make company's solutions data-driven.

Facebook, Amazon, Google and a lot of other big companies invest millions dollars into the NLP related studies to make their text- and speech-processing products really powerful and smart [25].

NLP tasks require strong mathematical skills, deep understanding of machine learning principles and proficiency in linguistics at once, so not every programmer can realize its approaches from scratch. That is why before NLP projects and researches were accessible only to high-level professionals. Luckily, nowadays various libraries allow developers with different background to use NLP algorithms – that makes language analysis simple as never before and gives developers a chance to focus on new challenges, especially ones connected with machine learning.

2.1.1 Python Tools

Python is multi-paradigm programming language (it supports functional, procedural and object-oriented programming). It considered to be general-purpose and high-level interpreted. It was designed by Guido van Rossum and released in 1991 [26]. It has very brief and readable syntax that includes using of whitespaces instead of semicolons. Its simplicity allows developers to write well-structured and clear code that can be easily scaled and will be suitable both for little and big programming projects [27]. Python has built-in garbage collection and is dynamically typed. Its standard library is very vast and rich. It supports a lot of NLP tools [28].

There are:

1. Natural Language Toolkit (NLTK)

Natural Language Toolkit (NLTK) is the most popular NLP tool for Python nowadays. It is also probably the most powerful and rich one as well. It has an implementation for each Natural Language Processing feature, and often – a few ones, so the developer has a choice what algorithm he will use for the task he has.

The NLTK library is a package of libraries and programs for the symbolic and statistical processing of a natural language. Contains graphical representations and sample data. It is accompanied by extensive documentation, including a book explaining the basic concepts behind the tasks of natural language processing that can be performed using this package.

NLTK is well-suited for students studying computer linguistics or related subjects such as empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning.

NLTK has been successfully used as a training tool, as an individual learning tool and as a platform for prototyping and creating research systems [29].

2. SpaCy

The main rival of NLTK is SpaCy. It works more quickly but has not a big variety of NLP elements realization. Its units are introduced as objects so it is easier to process them. Thus also makes this library compatible with a lot of other tools, particularly data science frameworks etc. so the developer can make deeper analysis of the text. The disadvantage of SpaCy is poorer choice of languages available. The interface is very brief and convenient. The documentation is very comprehensive. The library also offers different pre-built neural networks that cover a wide range of NLP and text analysis tasks. SpaCy is good for efficient apps that solve typical NLP problems [30].

3. TextBlob

TextBlob is like an NLTK add-in – a lot of Natural Language Toolkit tools are available while using TextBlob. The NLTK Pattern library functions are also accessible. This tool is awesome for little projects, it is also great for beginners – TextBlob gives a good insight into NLP field in a very convenient and simple format. This library can be called very efficient so it is better not to use it for the big projects [31].

4. Textacy

Textacy is a tool build on the base of SpaCy. It proposes the great pre- and post-processing functions – works with raw text to prepare it for further analysis. It can help to choose the right pipeline, show some statistics, compare documents and their parts etc.

It may appear an amazing option to use it if the developer decided to use SpiCy as his basic tool so it will be available to save some time on coding functions that are already realized in Textacy [32].

5. PyTorch-NLP

It is used to solve various problems: computer vision, natural language processing. Developed primarily by the artificial intelligence group Facebook. Also around this framework is built an ecosystem consisting of various libraries developed by third-party teams: Fast.ai, which simplifies the process of learning models, Pyro, a module for probabilistic programming, from Uber, Flair, for processing natural language and Catalyst, for training DL and RL models [33].

6. CoreNLP

A popular natural language processing toolkit that supports many of the core NLP tasks.

There are three supported ways to run CoreNLP tools:

- use of the base fully customizable API;
- using the Simple CoreNLP API;
- using a CoreNLP server.

7. Gensim

Finds similarities in two documents using vector and topic modeling. Library of natural language processing of destination for “Thematic modeling”.

With it, you can process texts, work with vector models of words (such as Word2Vec, FastText, etc.) and create thematic models of texts.

It is very efficient, optimized and it processes data really very fast thanks to NumPy [34].

8. polyglot

This slightly lesser-known library is one of our favorites because it offers a broad range of analysis and impressive language coverage. Thanks to NumPy, it also works really fast. Using polyglot is similar to spaCy – it’s very efficient, straightforward, and basically an excellent choice for projects involving a language spaCy doesn’t support. The library stands out from the crowd also because it requests the usage of a dedicated command in the command line through the pipeline mechanisms. Definitely worth a try [35].

9. scikit–learn

This handy NLP library provides developers with a wide range of algorithms for building machine learning models. It offers many functions for using the bag-of-words method of creating features to tackle text classification problems. The strength of this library is the intuitive classes methods. Also, scikit-learn has an excellent documentation that helps developers make the most of its features.

However, the library doesn’t use neural networks for text preprocessing. So if you’d like to carry out more complex preprocessing tasks like POS tagging for your text corpora, it’s better to use other NLP libraries and then return to scikit-learn for building your models [36].

10. Pattern

Another gem in the NLP libraries Python developers use to handle natural languages. Pattern allows part-of-speech tagging, sentiment analysis, vector space modeling, SVM, clustering, n-gram search, and WordNet. You can take advantage of a DOM parser, a web crawler, as well as some useful APIs like Twitter or Facebook. Still, the tool is essentially a web miner and might not be enough for completing other natural language processing tasks [37].

2.1.2 Node Tools

Node is a software platform based on the V8 engine (translating JavaScript into machine code) that turns JavaScript from a highly specialized language into a general-purpose language [38]. Node.js adds the ability for JavaScript to interact with input/output devices through its API (written in C ++), to connect other external libraries written in different languages, providing calls to them from JavaScript code [39]. Node.js is mainly used on the server, acting as a web server, but it is possible to develop desktop window applications on Node.js (using NW.js, AppJS or Electron for Linux, Windows and macOS) and even program microcontrollers (for example, tessel and espruino). Node.js is based on event-driven and asynchronous (or reactive) programming with non-blocking I/O [40].

1. Retext

Key features:

– full support for Markdown and reStructuredText, as well as Python-Markdown extensions;

– export to HTML, PDF, ODT from the box, as well as the ability to create your own export extensions (for example, there is an extension for uploading to Google Drive);

– tab support;

– support for CSS styles and syntax highlighting;

– spelling check;

– two viewing engines: one based on QTextBrowser and one based on WebKit [41].

2. Compromise

Compromise certainly isn't the most sophisticated tool. If you're looking for the most advanced algorithms or the most complete system, this probably isn't the right tool for you. However, if you want a performant tool that has a wide breadth of features and can function on the client side, you should take a look at Compromise. Overall, its name is accurate in that the creators compromised on functionality and accuracy by focusing on a small package with much more specific functionality that benefits from the user understanding more of the context surrounding the usage [42].

3. Natural

Natural includes most functions you might expect in a general NLP library. It is mostly focused on English, but some other languages have been contributed, and the community is open to additional contributions. It supports tokenizing, stemming, classification, phonetics, term frequency–inverse document frequency, WordNet, string similarity, and some inflections. It might be most comparable to NLTK, in that it tries to include everything in one package, but it is easier to use and isn't necessarily focused around research. Overall, this is a pretty full library, but it is still in active development and may require additional knowledge of underlying implementations to be fully effective [43].

4. Nlp.js

Nlp.js is built on top of several other NLP libraries, including Franc and Brain.js. It provides a nice interface into many components of NLP, like classification, sentiment analysis, stemming, named entity recognition, and natural language generation. It also supports quite a few languages, which is helpful if you plan to work in something other than English. Overall, this is a great general tool with a simplified interface into several other great tools. This will likely take you a long way in your applications before you need something more powerful or more flexible [44].

2.1.3 Java Tools

Java is a general-purpose programming language that is class-based, object-oriented, and designed to have as few implementation dependencies as possible. It is intended to let application developers write once, run anywhere (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation.Java applications are typically compiled to bytecode that can run on any Java virtual machine (JVM) regardless of the underlying computer architecture [46]. The syntax of Java is similar to C and C++, but it has fewer low-level facilities than either of them. As of 2019, Java was one of the most popular programming languages in use according to GitHub, particularly for client-server web applications, with a reported 9 million developers [47].

Java was originally developed by James Gosling at Sun Microsystems (which has since been acquired by Oracle) and released in 1995 as a core component of Sun Microsystems' Java platform. The original and reference implementation Java compilers, virtual machines, and class libraries were originally released by Sun under proprietary licenses. As of May 2007, in compliance with the specifications of the Java Community Process, Sun had relicensed most of its Java technologies under the GNU General Public License. Meanwhile, others have developed alternative implementations of these Sun technologies, such as the GNU Compiler for Java (bytecode compiler), GNU Classpath (standard libraries), and IcedTea-Web (browser plugin for applets) [48].

The latest versions are Java 13, released in September 2019, and Java 11, a currently supported long-term support (LTS) version, released on September 25, 2018; Oracle released for the legacy Java 8 LTS the last free public update in January 2019 for commercial use, while it will otherwise still support Java 8 with public updates for personal use up to at least December 2020. Oracle (and others) highly recommend uninstalling older versions of Java because of serious risks due to unresolved security issues. Since Java 9 (and 10 and 12) is no longer supported, Oracle advises its users to immediately transition to the latest version (currently Java 13) or an LTS release [49].

1. OpenNLP

OpenNLP is hosted by the Apache Foundation, so it's easy to integrate it into other Apache projects, like Apache Flink, Apache NiFi, and Apache Spark.

It is a general NLP tool that covers all the common processing components of NLP, and it can be used from the command line or within an application as a library. It also has wide support for multiple languages. Overall, OpenNLP is a powerful tool with a lot of features and ready for production workloads if you're using Java [50].

2. StanfordNLP

A popular natural language processing toolkit that supports many of the core NLP tasks.

There are three supported ways to run CoreNLP tools:

- use of the base fully customizable API;
- using the Simple CoreNLP API;
- using a CoreNLP server [51].

3. CogCompNLP

CogCompNLP, developed by the University of Illinois, also has a Python library with similar functionality. It can be used to process text, either locally or on remote systems, which can remove a tremendous burden from your local device. It provides processing functions such as tokenization, part-of-speech tagging, chunking, named-entity tagging, lemmatization, dependency and constituency parsing, and semantic role labeling. Overall, this is a great tool for research, and it has a lot of components that you can explore. I'm not sure it's great for production workloads, but it's worth trying if you plan to use Java [52].

4. NLP4J

Natural Language Processing for JVM languages (NLP4J) project provides NLP tools readily available for research in various disciplines, frameworks for fast development of efficient and robust NLP components, API for manipulating computational structures in NLP (e.g., dependency graph). The project is initiated and currently led by the Emory NLP research group and is under the Apache 2 license [53].

5. Apache UIMA

Unstructured Information Management Applications (UIMA) is a component architecture and software framework implementation for the analysis of unstructured content like text, audio, and video data. The goal of UIMA is to transform unstructured information to structured information by orchestrating analysis engines to detect entities or relations and thus to build the bridge between the unstructured and the structured world. UIMA additionally provides capabilities to wrap components as network services and can scale to very large volumes by replicating processing pipelines over a cluster of networked nodes [54].

6. GATE Embedded

General Architecture for Text Engineering (GATE) is an open source software toolkit which is capable of solving almost any text processing system. GATE Embedded is an object-oriented open source framework (or class library) implemented in Java which is used in all GATE-based systems and forms the core elements of GATE developer. It is designed to allow you to embed language processing functionality in diverse applications [55].

7. LingPipe

LingPipe is a toolkit for processing text using computational linguistics. The tasks such as finding the names of people, organizations or locations in news, automatically classify Twitter search results into categories, etc. can be easily done by this toolkit. The architecture of LingPipe is designed to be efficient, scalable, reusable and robust. It has Java API with source code and unit tests, n-best output with statistical confidence estimates, thread-safe models and decoders for concurrent-read exclusive-write (CREW) synchronization, etc.

For our project development we choose Python. It has a wide range of the developed toolkits and libraries thus work with NLP gets much more simple and accessible. Among Python tools we are choosing TextBlob since it is functional and easy to use [56].

2.2 Selection of Target Social Network

By Wikipedia definition, social networking service (virtual social network - hereinafter referred to as "social network") is an online platform that people use to build social connections, interact with other people with shared careers or personal interests, views, experiences, likes, pursuits, or connections in real life .

The diversity of social networks is huge, but there are some hallmarks of all resources in this area:

– these are web 2.0 standard online applications;

– the content (content) of the web application data is generated by the users themselves;

– users create a resource-specific profile (account);

– social networking services deliver their development by linking user profiles to other users 'or groups' accounts.

According to a survey conducted by statista.com in September 2017, the following social networks are the most popular in the world:

1. Facebook – over 2 billion users;
2. YouTube has 1.5 billion users;
3. WhatsApp – 1.3 billion users;
4. WeChat – 963 million users;
5. QQ – 850 million users;
6. Instagram – 700 million users;
7. QZone – 606 million users;
8. Tumblr – 368 million users;
9. Sina Weibo – 361 million users;
10. Twitter – 328 million users;
11. Baidu Tieba – 300 million users;

12. Skype – 300 million users;

13. Viber – 260 million users [57].

Research on smartinsight.com in the US shows the following statistics:

1. By spread (% of population using the resource):

- Facebook – 63%;
- FB Messenger – 47%;
- Instagram – 27%;
- Twitter – 22%;
- Google Plus – 17%.

2. By the number of days per month when a person enters the social network:

- Facebook – 15.0;
- FB Messenger – 7.9;
- Instagram – 11.0;
- Twitter – 7.5;
- Google Plus – 3.2.

3. By number of sessions (logins) per day:

- Facebook – 8;
- FB Messenger – 3;
- Instagram – 6;
- Twitter – 5;
- Google Plus – 2.

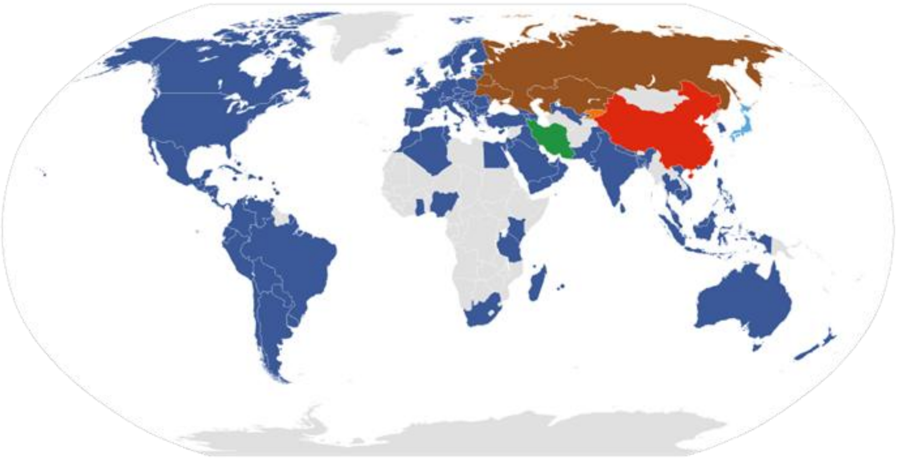


Figure 1.2 – Top social networking sites by country (dark-blue marked Facebook; blue Twitter; brown VKontakte; red QZone; orange Odnoklassniki; green Facenama; gray – non-reported countries) [57]

76% of Facebook users use the social network daily, 15% – are active during the week, 7% are active less frequently. The closest competitors to this indicator have the following results:

Instagram: 51% – 26% – 22%;

Twitter: 42% – 24% – 33% [58].

Let's review the concept of these three social networks.

2.2.1 Instagram

Instagram – a photo social networking site for photo hosting, connected to the application with the same name (available for Windows Phone, iOS and Android operating systems), which allows users to edit photos and videos, apply digital filters to them, and share them on various sites social networks. A characteristic feature of the application was the square shape of the photos, similar to Kodak Instamatic cameras, Polaroid instant cameras and 6x6 medium format cameras as opposed to the 4:3 aspect ratio, which is used by most cameras and mobile devices with photo function. On August 27, 2015, the option of placing photos and videos in other image formats was made available.

In April 2012, the website was bought by Facebook for about USD 1 billion (only 13 employees were working on it at the time) [59]. After some time, the regulations were changed – new regulations indicate the possibility of selling user photos to external companies. It is estimated that this was one of the main reasons for the decrease in the number of active users - while in September 2012 about 100 million users used the application and website, in January 2013 there were about 90 million. In September 2017, Instagram had 800 million active users per month, including 500 million visitors every day. Instagram's constant problem is fake accounts used to generate spam on the platform. Despite the successive removal by administrators, at the turn of June and July 2015 the number of inauthentic profiles was estimated at over 20 million spambots.

On December 9, 2012, Instagram turned off support for photos published on Twitter (from that day it is not possible to display photos taken using Instagram on Twitter) [60].

In the first week of July 2015, the photo upload began to Instagram servers with a resolution of 1080x1080. It was the first stage of replacing the standard 640x640 format, present since the beginning of the platform's operation. Instagram did not announce the transition to 1080 pixels in the official announcement. The change in the quality of photos was noticed by users in the source code of the page and concerned only the mobile application [61].

On July 20, 2015, a search engine was introduced on the Internet version of Instagram, enabling search of hashtags (#), user profiles and locations. The search options also include geotags and an overview of the most popular posts. Change of search mechanisms was carried out successively on individual markets.

Currently, the largest number of followers have profiles of pop culture and sports stars such as Cristiano Ronaldo, Selena Gomez, Ariana Grande, Dwayne Johnson, Kim Kardashian West.

1. The Graph API for Instagram allows Instagram Business users and authors to manage their accounts through your app. With its help, you can receive and publish media files, manage and reply to comments, find media files in which the user was @mentioned, find media files with hashtags and receive basic metadata and metrics of other Instagram Business users and Instagram authors.

This API is intended for Instagram Business users and authors who need statistics on their interactions on social networks and their management. If your application is intended for ordinary users or for obtaining only basic information about the user profile and the photos and videos published by him, use the Instagram Basic Display API.

2. Mentions API

Simplify the interaction between people and companies. If a person sees that he is marked on the photo or mentioned with the @ symbol, he can immediately write back.

3. Business Discovery API

Track the actions of competitors. Browse organic posts from other companies to see how they promote themselves on Instagram.

4. API Insights

Help companies track and analyze the metrics of their business profile on Instagram. Offer them tools to optimize the performance of organic materials on Instagram.

5. API Comment Moderation

Communication without borders. Help companies communicate effectively with their Instagram audience through comments [62].

2.2.2 Facebook

Facebook is a social network and social media from America that has been founded by Harvard students Mark Zuckerberg, Andrew McCollum, Eduardo Saverin, Chris Hughes, and Dustin Moskovitz.

First, the network worked only for students of a few universities, but in 2006 the registration on the website has become available for all people that confirmed that they are above 13 years old. Face book – handbooks that are given to students in some universities in the USA.

Users can get access to Facebook from various devices – computers, laptops, smartphones, etc. The only condition is an Internet connection. To start using Facebook user has to register the profile. He or she can add a profile photo, personal information, friends, post text, photos, and videos, sharing this data with the public, friends, some group of friends or one particular friend – depending on what privacy settings have been set. A person is considered to be somebody's friend after one of them sent a request and others accepted it.

There is also a big variety of groups, public pages, events, videos, embedded applications, and games. Users can be invited to an event, to like some page, receive notifications about events, activities and birthdays of his or her friends, comment posts on personal profiles of public pages.

As of December 2018, there were almost 2.5 billion monthly active users. But there also exists a big problem with the fake accounts – at the end of 2018-beginning of 2019 around 3 million fake accounts had been detected. It is unknown how many fake accounts remained uncaught, so it may appear hard to estimate the real number of users precisely [63].

Facebook draws a huge mass media attention. A few times Facebook got to the epicenter of the scandals, like the situation with the users' data leakage (Cambridge Analytica scandal), with the U. S. elections in 2016, etc. There are also controversies with fake news, inappropriate content, copyright violation, social media addiction, unreliable advertisement, etc. It is often blamed in serving as a platform for fake news spreading, so in 2017 Facebook decided to cooperate with Poynter Institute's International Fact-Checking Network in order to detect fakes. Nevertheless, most political advertising is skipped, so Facebook is still being accused of weak false data control [64].

Facebook has a tool for third-party applications that allows them to access users' information, so-called social graph – the name of the tool is Graph API. All Facebook products, APIs and SDKs have connections to the Graph API, so it is a very important tool when it comes to the work with this social media.

The Graph API allows observing users' data in the unified form, as objects of a graph. They may consist of images, humans, posts, events, pages, friends' nets, connections between people and other listed things. To get access to private information developer has to obtain a special key. It is not needed when it comes to public data. Graph API helps to design functional applications that can be monetized [65].

To make Graph API a part of an application developer needs to have a special access token, which is possible to get during application registration. While an application calls Graph API, the request goes through the graph.facebook.com or other host URL. It reaches the node that the developer expects to attain via object ID [66]. GET request helps to get any node/edge. It is also needed to add relevant parameters. There is also an option to sort the data by time.

2.2.3 Twitter

Twitter is a social network for public messaging using the web interface, SMS, instant messenger or third-party client programs for Internet users of any age. The publication of short notes in a blog format is called “microblogging”. Using the service is free. Use via SMS is charged by the operator according to the user's tariff plan [67].

Twitter is owned by Twitter Inc., with its headquarters in San Francisco. As of June 2015, about 3,000 employees work in the company's American offices [68].

Created by Jack Dorsey in 2006, Twitter soon gained worldwide popularity. As of January 1, 2011, the service has more than 200 million users. 100 million users are active at least once a month, of which 50 million use Twitter every day. 55% use Twitter on mobile gadgets, and about 400 million unique visits are received per month [69].

In 2014, the revenue of Twitter Inc. amounted to 1.4 billion dollars [70].

One of the access interfaces to Twitter is the web API. This API allows, among others for tracking live user activity, publishing new tweets or searching them.

Thanks to the API, it is possible to build your own applications / services using Twitter functions, without the need for a standard web GUI. In some cases, the application interface consists of a single button (allowing, for example, to place specific content in the form of a tweet) [71]. The implementation of such functions is relatively simple and requires the use of a limited part of the Twitter API functionality that performs the function needed. According to Kevin Makice, the author of the Twitter API book, the creation of the Twitter web API preceded the creation of Twitter applications, and in 2008 there were more than 1,000 applications using this API.

An example of using the API to contact users is the NBA sports organization, which publishes on Twitter information about sporting events and plans, results, as well as encloses statements of activists and athletes [72].

Because Twitter posts are a rich source of content, the Twitter API has become a popular research tool, allowing for relatively simple statistical analysis of a large number of tweets. For example, in one study (which concerned the possibility of detecting an influenza epidemic based on Twitter content analysis), over 300 million tweets were analyzed using the API. Another study carried out a continuous analysis of selected channels (over 30,000 per day) to build an early detection system for hazardous events [73].

2.3 Choosing Bot Platform

There are various platforms for bot development – literally every powerful social networks allows to create bot on it basis using its own API. We will review the most famous ones.

2.3.1 Facebook Messenger

More than 1.3 billion peopleuse the Facebook Messenger chatbot platform every month.It is considered one of the greatest and most powerfultools for programmers. The Facebook chatbotsare basedon personal chatting with a public page. The initiator is always the user. Two main things the developerhas to do to create a botin Facebook are: to own a company page in public accessand create an application to obtain access to the API [74].

Facebook chatbot has a capability to interplay with the user both through regular text messages and "Structured Text"(specific markup language). This languagesupports the usage of buttons, items, and invoices in correspondence.

1. Buttons

Serve to send messages that require user response. Up to three buttons can be nestedin a single message.

Facebook API supports three kinds of buttons:

– ones that send a response to a bot;

– link buttons – they direct to some URL on the Internetafter hitting;

– forward messages to the responsible person, i.e. manager (in cases when additional help or assistance of the real person is needed).

2. Elements

Designed to send product profile or anything of an alike view. Such element can contain a heading, subheading, characteristics, picture, and buttons. The message may contain up to ten items.User can scroll them horizontally.

3. Invoices (paid)

They give the chatbot apossibility to operate as a shopbot. While the person clicks button "Buy", the bot generates an account that containsthe whole data about the purchase with the possibility to payright away. The feature currently works in the test mode [75].

The bot in Facebook is not granted with the usual user rights. It is not capable of initiating a dialogue on its own with either subscribers or random users. To make bot work one has to open it and start a conversation.

2.3.2 Viber

To create Viber bot developer should have a public account, usually the oneorganization uses to communicate with its customers. This account can be managed both by a person and a bot, that will interact with the users on behalf of the organization. Despite this Viber bot is not inferior to bots on other platforms. It has all the standard functions, including the buttons, bills, and purchases.

Other specialties Viber proposes:

– chatbotsin Viber canstart a conversation with any contact, so the user should not be its subscriberto receive a message. The only condition is to have customers' phone numbers and Viber application installed on their devices. Additionally, unlike in the case of using regular SMS and emails, in Viber users can send a feedback to the messages they receive in distro [76];

– Viber botshave a capability to send carousel messages. They containitems (commodities) that are placed one by one horizontally so user can scroll them in a circle. Everyelement includes an item image, characteristics, cost and a button "Buy". Such format is practical andconvenient for the user. It isvery common inadvertisement on sites in the Interner, and can also be used in a messenger.

Admin panel for bots in Viber is very plain and easy to use. It gives owner an opportunity to manage all his public accounts, view statistics and pages' information. It also allows to make a bot right in Viber and easily adjust its work without any additional tools. This bot will be equipped with the basic settings and functions and may work not as expected.

2.3.3Telegram

One of the first chatbots appeared in Telegram. Unlike other messengers, Telegram immediately shows that you are communicating with a bot, because the name of chatbot always ends with the word "bot" – it is the key requirement. Control of the chatbot is quite convenient: for this purpose, a special bot was created – its name is BotFather. BotFather manages all the bots, it is used to create Telegram bots, to control them and their functionality.

Also, Telegram bots allow to proceed payments.

Telegram bots can:

– perform typical chatbot features (translate, comment, teach, test, search for information, play and entertain, play video and audio, make payments and other bank operations, contain embedded games, and more);

– execute additional API-accessible bot functions (integrate with other services and platforms, interact with devices and other user’s things, that are connected to the Internet) [77].

There are "standard" chatbots in Telegram as well. These are helpers that check grammar, conduct surveys, tell anecdotes and send reminders of meetings and chores. They can simultaneously gather statistics and keep track of company chat activity. You can add them to your chats or use their functionality to create your own bot. For example, the Botan analytics system contains Yandex application metrics integrated with the Telegram bot.

Telegram bot capabilities include the inline built-in mode. This mode allows users to access the bot from any chat. In any Telegram chat, you can enter a message with the name of the bot, and then the request itself. The bot activates after mentioning its name and responds to the request directly in open chat. You can also add a "Switch to PM" button to the built-in bot. If you click on it, the bot will open in a new chat.

Conveniently, the inline mode can be activated for any bot, both in process of creating a new one and for already existing.

2.3.4Skype

Following the trend, Microsoft has made the Skype bot platform, and developers got an opportunity to create bots for Skype. It is stated that in the future it will be possible to create audio and video bots in addition to text ones. For video calls, various graphically drawn characters will play the role of the interlocutor. Following the Telegram bot principle, Skype also created some standard bots that tell you the latest news, search for music tracks and photos.

Microsoft has made a loud statement that with the help of bots, they plan to change the shopping, food order, travel arrangements processes. These are just big plans right now. Currently, Skype proposes existing text bots and creating your own text bots on the Skype bot platform.

The Skype chatbot has the same message elements as the others (buttons, items, and forms for payment). But the functionality is still partially in test mode and has some inconveniences. There is an interesting “Menu” feature to greet the user, which platform developers recommend to use for search robots. Indeed, opening the bot, it is convenient to immediately see a list of its capabilities.

Much attention has been paid to the language that bots use, especially given Microsoft's plans to develop talking bots.

Skype bot Cognitive Service Language APIs allow the bot to:

– collect keywords from the text and build on the assumption of queries;

– build assumptions based on context and experience (requests that have previously come from this or other users);

– independently generate complete spoken responses from a semi-structured set of response templates.

This system is already one step closer to human thinking, as it is capable of learning and making assumptions using its experience.

A Skype chatbot can integrate with Google Analytics and pull in the required data from there, with the ability to customize analytics for a specific user (select options and format of results).

As for the bot's ability to initiate a dialogue with a client, Skype cannot send messages to users who have not been added to the contact list. Therefore, you cannot fully rely on the bot to send messages on Skype [78].

The biggest advantage of Skype as a chatbot platform is the fact that its functionality is constantly updated, and Microsoft has "big plans" for it. Of course, there are also unpleasant moments – it has not yet been thoroughly tested, and while creating/registering bots developer often faces all kinds of problems.

2.3.5VKontakte

VKontakte (VK) bot lives on a public page – community, group, meeting. The logic of the bot will be provided by a script in which you can lay different event scenarios. For example, VK bot can not only communicate with users, but also monitor activity on the page and give the corresponding reactions to it. There are two tools for this: Callback API and Long Poll.

Callback API can notify about events in community by sending information to the server as soon as an event occurs. Long Poll does the same, but just after a request of an administrator, and displays the data about new events all in one list. Callback API notifies about any events: wall posts, new user requests, comments and more. Long Poll will only report events related to messages. When using both tools, messages with complete information will be sent to your server in JSON format. Developer can also add a script of reaction to these events to VK chatbot.

Vkontakte has its own payment API for chatbots, but only for virtual money and payments for virtual products in different applications. Therefore, VK bot will not replace the online store completely. But despite this, VK is more popular for shopping than any other social network [79].

The most used VK bots are: ones that work VK groups, ones that send messages Vkontakte, spam detecting ones.

2.3.6 Slack

Slack is an awesomebusiness chat application. It iswidely usedby teamsfor communication among the members and corporate correspondence. Using Slack platform, a developer can produce bots of 2kinds: application and user ones. An application bot has an ability to operate not only in the messenger, but in other parts of program as well.The other kind of bots can operate justin the messenger and can be added to different chats [80].

The first kind helpsto communicate; it is designed to assist in the intrinsiccircles of the organization. The second is a segregate mobile application that has a Slack interface and contains an integration tomessenger.Some useful Slack bots have been created to demonstrate the full potential of the platform. Major partis aimed to ease and simplify the job of organizations’managers as well as automate it. Examples:

– Relay and Birdly – trackuser’s actions;

– Nikabot – surveys staff on task completion and create reports about the current state of work;

– Meekan – organizingbot: helps to plan meetings, considering the schedule of everymember, chooses a comfortable time for everybody and schedules an appointment.

3 DEVELOPMENT AND TESTING OF AN APPLICATION

3.1 Work with Twitter API

To start work with Twitter API we have to register in Twitter in order to apply for a developer account. Next step is to create an application. Since in the recent year the process of Twitter application registration has become more complicated (the developer has to write quite a long and descriptive text with the reasoning why he needs to access Twitter data), we will use an old application (figure 3.1) that has been created in 2018 within the confines of writing the bachelor thesis – Twitter allowed old applications not to reregister. The only thing we need from it are keys and tokens, so it suits perfectly. There are four credentials we need to access Twitter: API key, API secret key, access token, access token secret. This data is confidential and only developer should have an access to it.

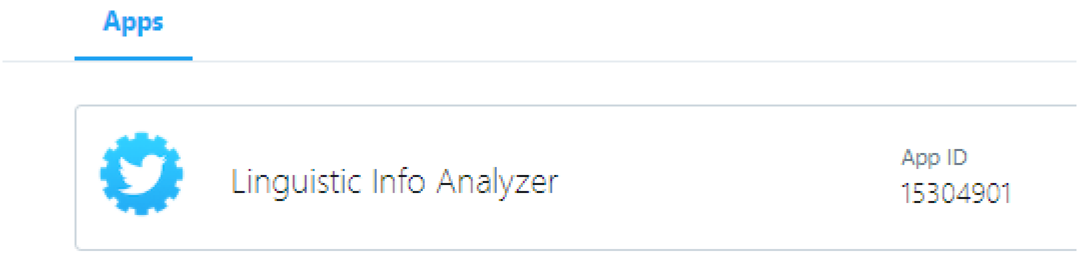
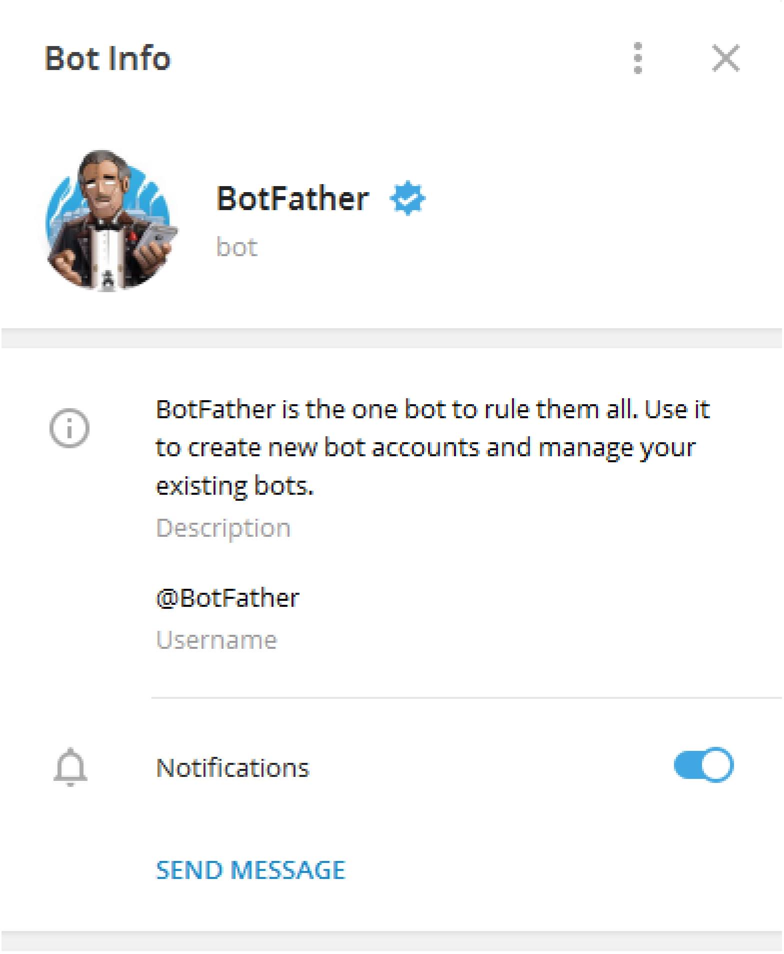


Figure 3.1 – Our Applicationin Twitter

For more convenient work with Twitter API we will use a special Python library – Tweepy, it works well with the tweets, longer than 140 characters and is brief and reliable.

3.2 Interaction with Telegram Bot API

As we mentioned before, Telegram has a bot that is a boss of all other bots. Its name is BotFather. BotFather is a very handy tool that simplifies and standardizes the creation of apps of this type (figure 3.2). To create a bot, we will send him command /newbot. After that bots offers us to write a name and username for our bot (figure 3.3). The main conditions – it must end with the word “bot” so the users will always understand that they are chatting with it, not with the human-being.

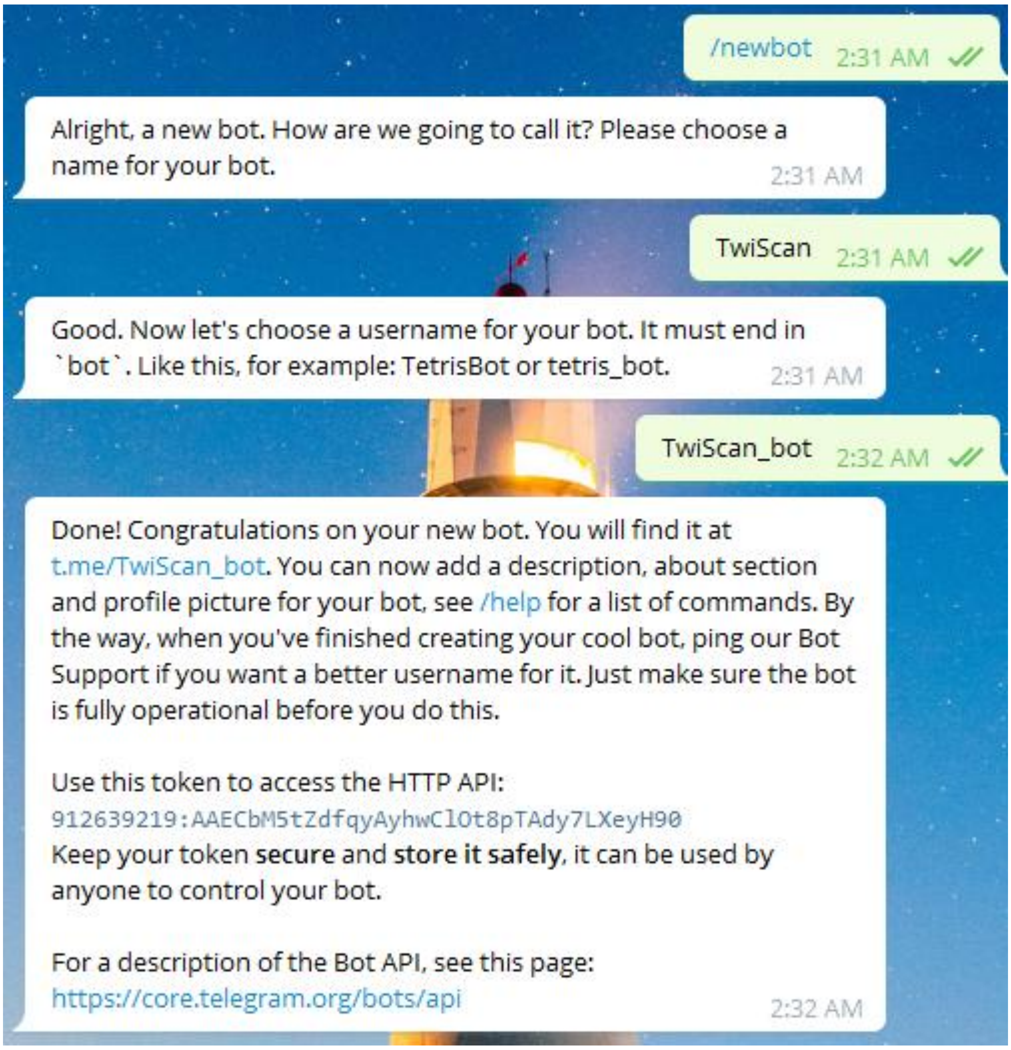


BotFather
bot

BotFather is the one bot to rule them all. Use it to create new bot accounts and manage your existing bots.
Description

@BotFather
Username

Figure 3.2 – BotFather in Telegram



/newbot 2:31 AM

Alright, a new bot. How are we going to call it? Please choose a name for your bot. 2:31 AM

TwiScan 2:31 AM

Good. Now let's choose a username for your bot. It must end in `bot`. Like this, for example: TetrisBot or tetris_bot. 2:31 AM

TwiScan_bot 2:32 AM

Done! Congratulations on your new bot. You will find it at t.me/TwiScan_bot. You can now add a description, about section and profile picture for your bot, see /help for a list of commands. By the way, when you've finished creating your cool bot, ping our Bot Support if you want a better username for it. Just make sure the bot is fully operational before you do this.

Use this token to access the HTTP API:
`912639219:AAECbM5tZdfqyAyhwCl0t8pTAdy7LXeyH90`
Keep your token **secure** and **store it safely**, it can be used by anyone to control your bot.

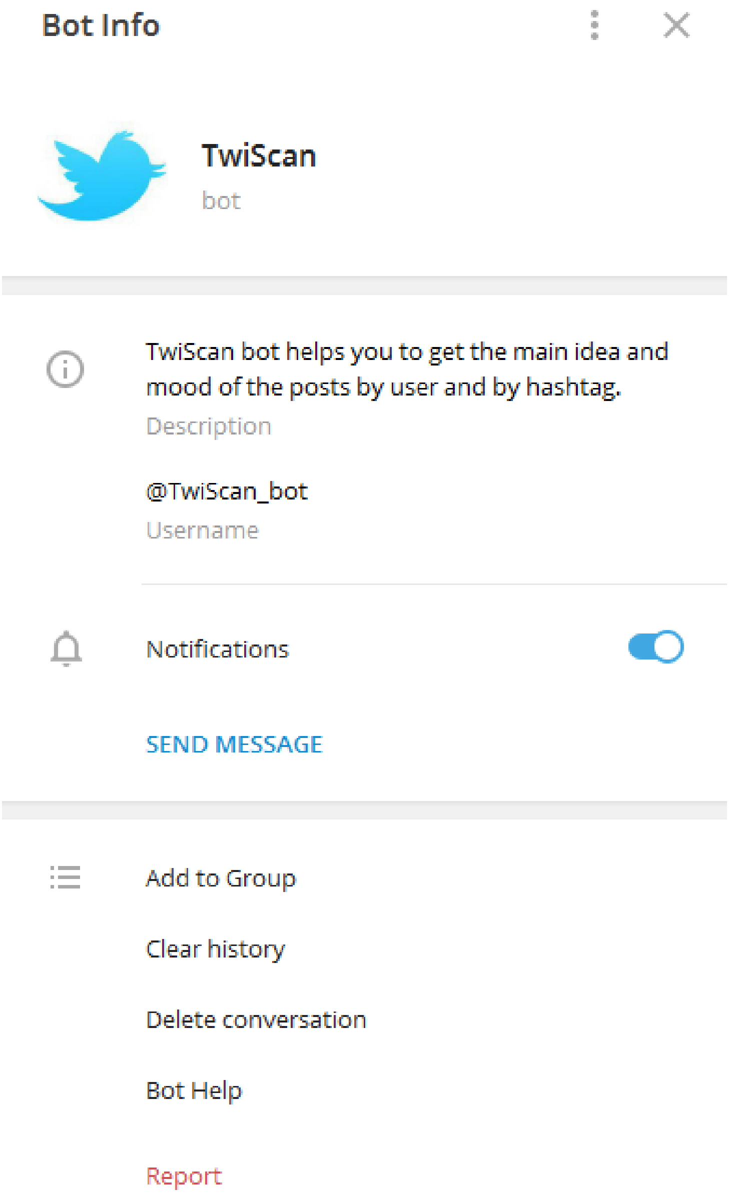
For a description of the Bot API, see this page:
https://core.telegram.org/bots/api 2:32 AM

Figure 3.3 – The process of bot creation

After bot registration we get access token that is secret and may is used for bot control. We also added a description – users will see it in the bot info window (figure 3.4) and about information that will be visible while we just start our work the bot (figure 3.5). Telegram does not limit the developers very much, but kindly asks them to implement the standard command – /start and /help (as you may have noticed, all the commands in Telegram start with the slash). We implemented them in our bot as well (figure 3.6). It is done with the purpose to make bot usage intuitive and convenient.

3.3 Application Functionality Overview

When we firstopen the bot, it gives us a brief explanation about its functions, as it is shownon the figure 3.5. To start bot user has to call /start command. After that bot asks how would user like to search the posts for analysis. There are two options: by hashtag or some keyword or by username. These options are represented as keyboard buttons.



Bot Info

TwiScan

bot

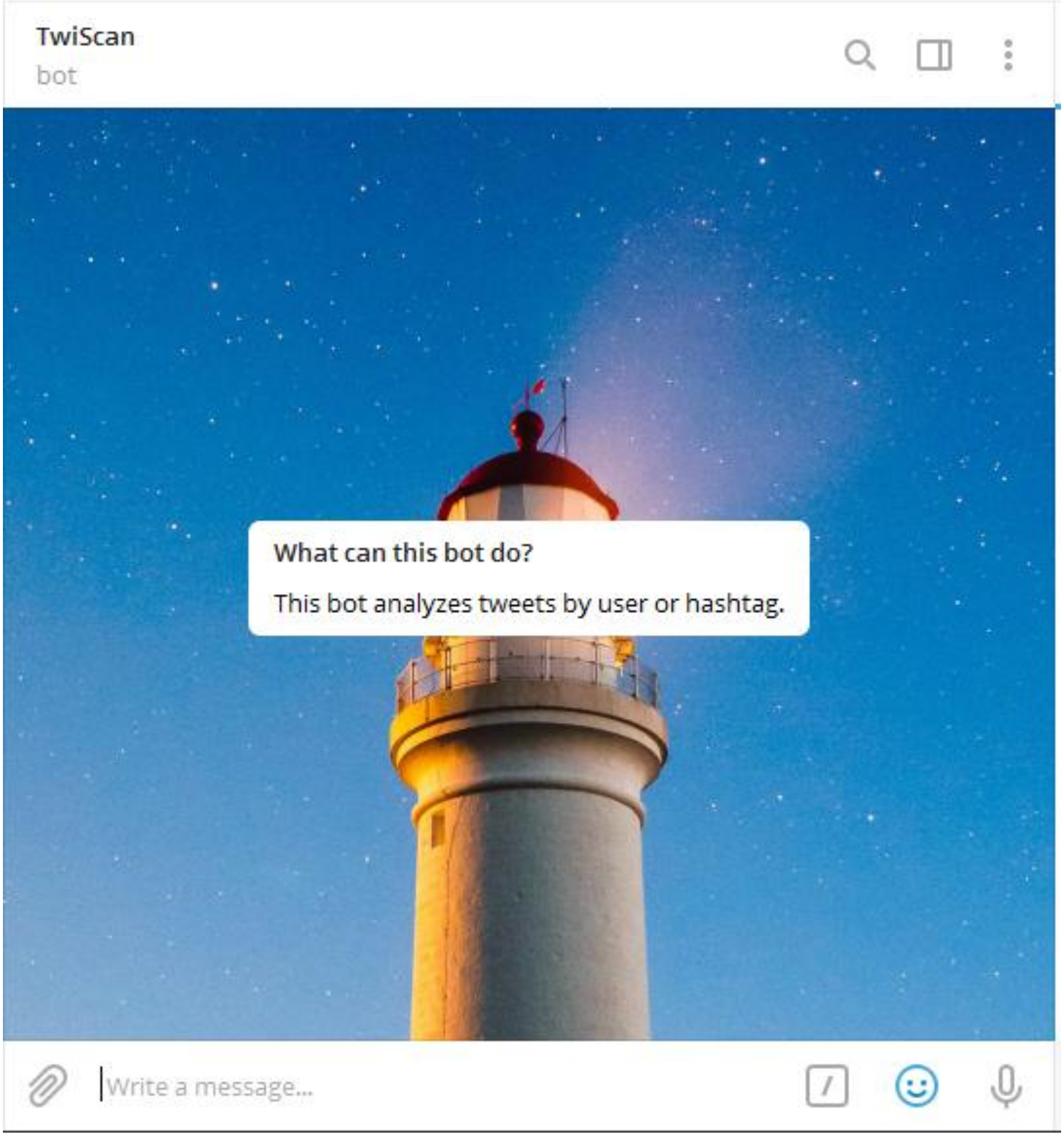
TwiScan bot helps you to get the main idea and mood of the posts by user and by hashtag.

Description

@TwiScan_bot

Username

Figure 3.4 – Bot Info



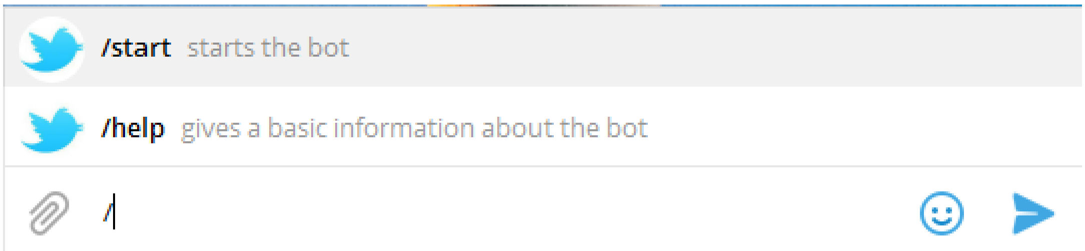
TwiScan

bot

What can this bot do?

This bot analyzes tweets by user or hashtag.

Figure 3.5 – Start of the interaction with bot

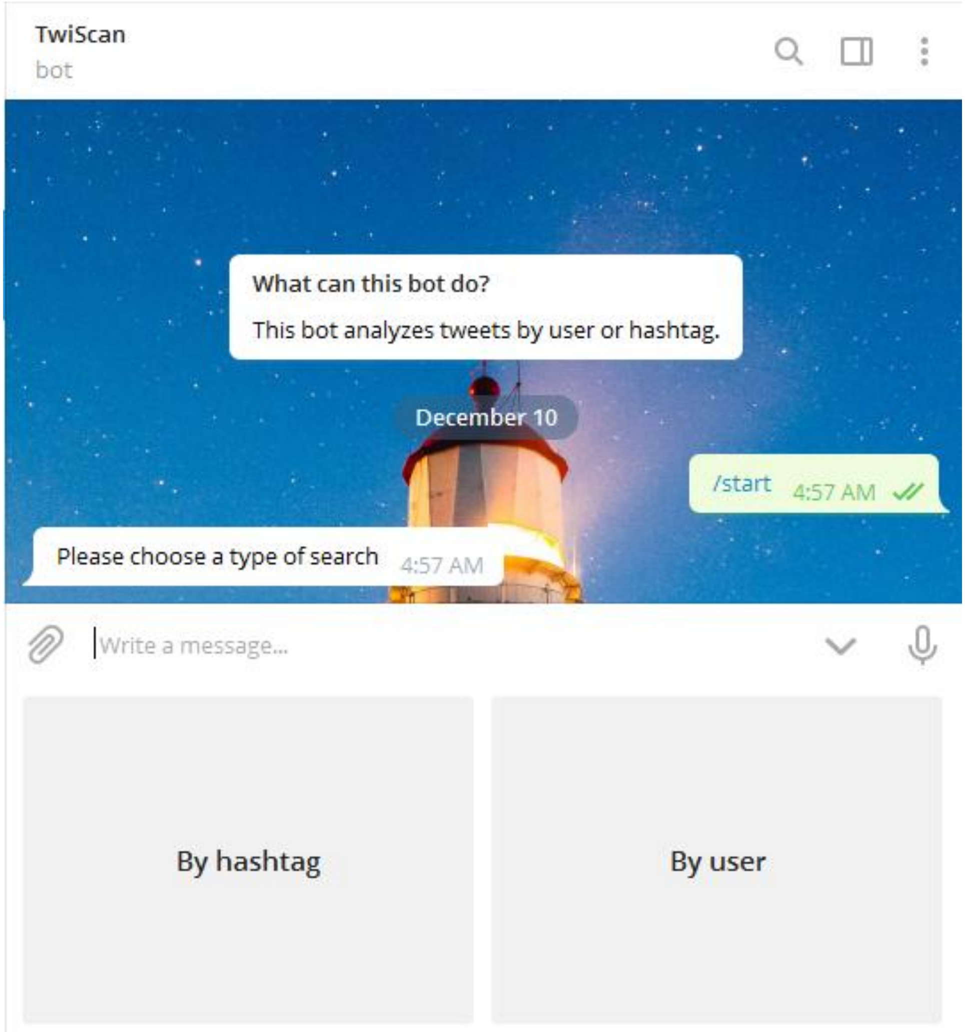


/start starts the bot

/help gives a basic information about the bot

/

Figure 3.6 – Standard commands implemented in our bot



What can this bot do?

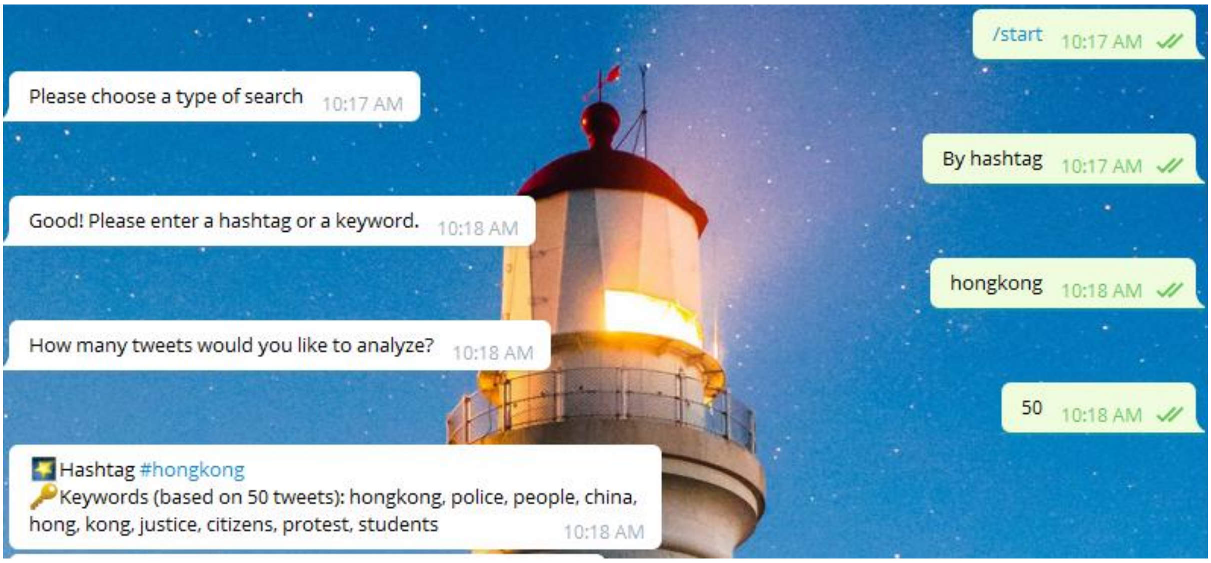
This bot analyzes tweets by user or hashtag.

December 10

/start 4:57 AM

Please choose a type of search 4:57 AM

Figure 3.7 – Choosing a type of search



/start 10:17 AM

Please choose a type of search 10:17 AM

By hashtag 10:17 AM

Good! Please enter a hashtag or a keyword. 10:18 AM

hongkong 10:18 AM

How many tweets would you like to analyze? 10:18 AM

50 10:18 AM

🌟Hashtag #hongkong
🔑Keywords (based on 50 tweets): hongkong, police, people, china, hong, kong, justice, citizens, protest, students 10:18 AM

Figure 3.8 – Keywords analysis of tweets that contain hashtag “hongkong”

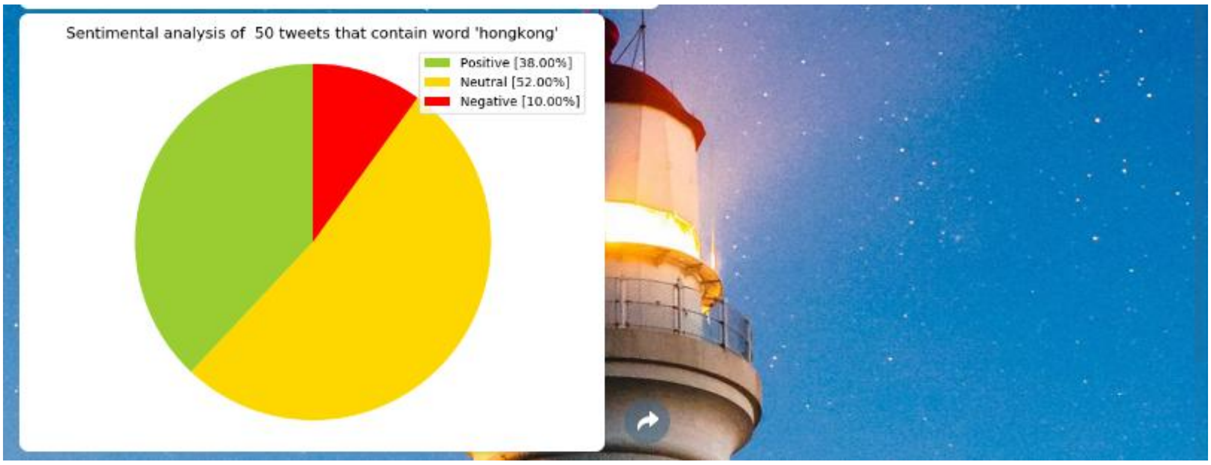
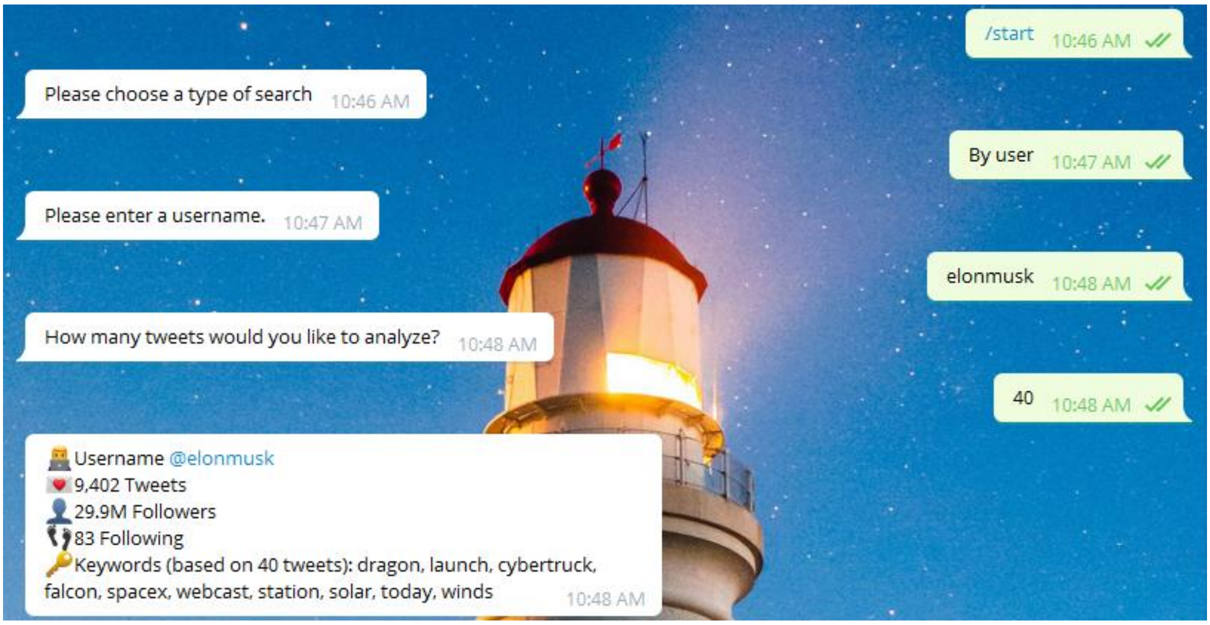


Figure 3.9 – Sentimental analysis of tweets that contain hashtag “hongkong”



/start 10:46 AM

Please choose a type of search 10:46 AM

By user 10:47 AM

Please enter a username. 10:47 AM

elonmusk 10:48 AM

How many tweets would you like to analyze? 10:48 AM

40 10:48 AM

Username @elonmusk
9,402 Tweets
29.9M Followers
83 Following
Keywords (based on 40 tweets): dragon, launch, cybertruck, falcon, spacex, webcast, station, solar, today, winds 10:48 AM

Figure 3.10 – Keywords analysis of tweets of user with the username “elonmusk”

After choosing a single option user has to input a number of tweets that must be analyzed. Bot represents a result of keyboard analysis (the quantity of keywords is limited to 10, because usually 10 keywords are enough to get the meaning of idea of text). It also shows us sentimental analysis of tweets. The keyboard and sentimental analysis is done on the basis of TextBlob tools. The diagram is built with the use of matplotlib – a library for data visualizing. In case if the search is done by username, we also get general information about his profile. It is important to remember that Twitter has daily limits, so it is impossible to use app if the limit has been reached.

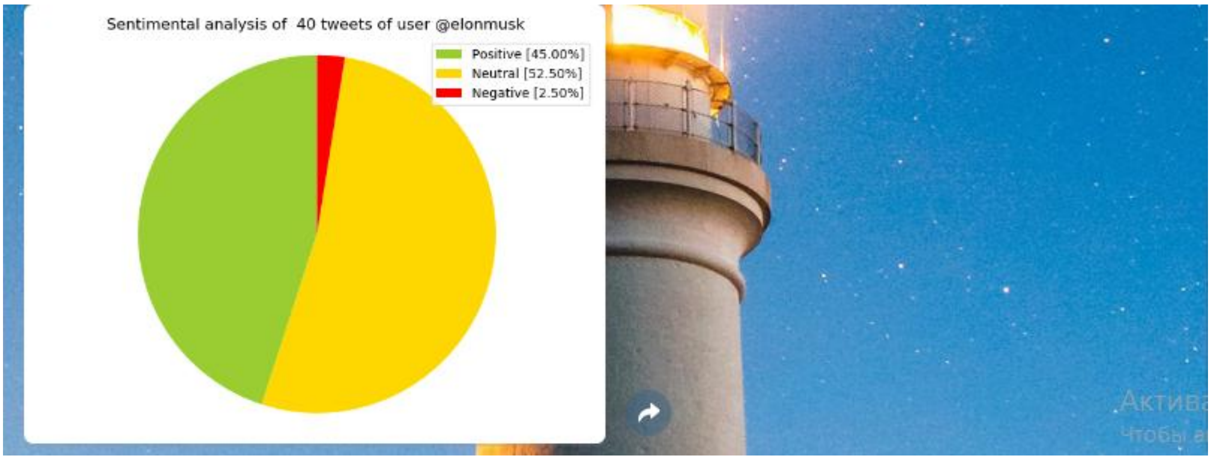


Figure 3.11 – Sentimental analysis of tweets of user with the username “elonmusk”

3.3 Analysis of the Obtained Results

After analysis of 50 tweets with the hashtag “hongkong” our application determined the following keywords:hongkong, police, people, china, hong, kong, justice, citizens, protest, students.

Let’s compare obtained results with the results of the websites with analogical tools available. We will test them on the same set of tweets.

http://textalyser.net/ extracts keywords, determines occurrences, frequency, and ranks keywords. Top-10 keywords (with the occurrences, frequency pointed in the brackets):

1. hongkong (33, 4.7%);
2. kong (9, 1.3%);

3. hong(9, 1.3%);

4. police (9, 1.3%);

5. people (8, 1.1%)

6. them (6, 0.8%);

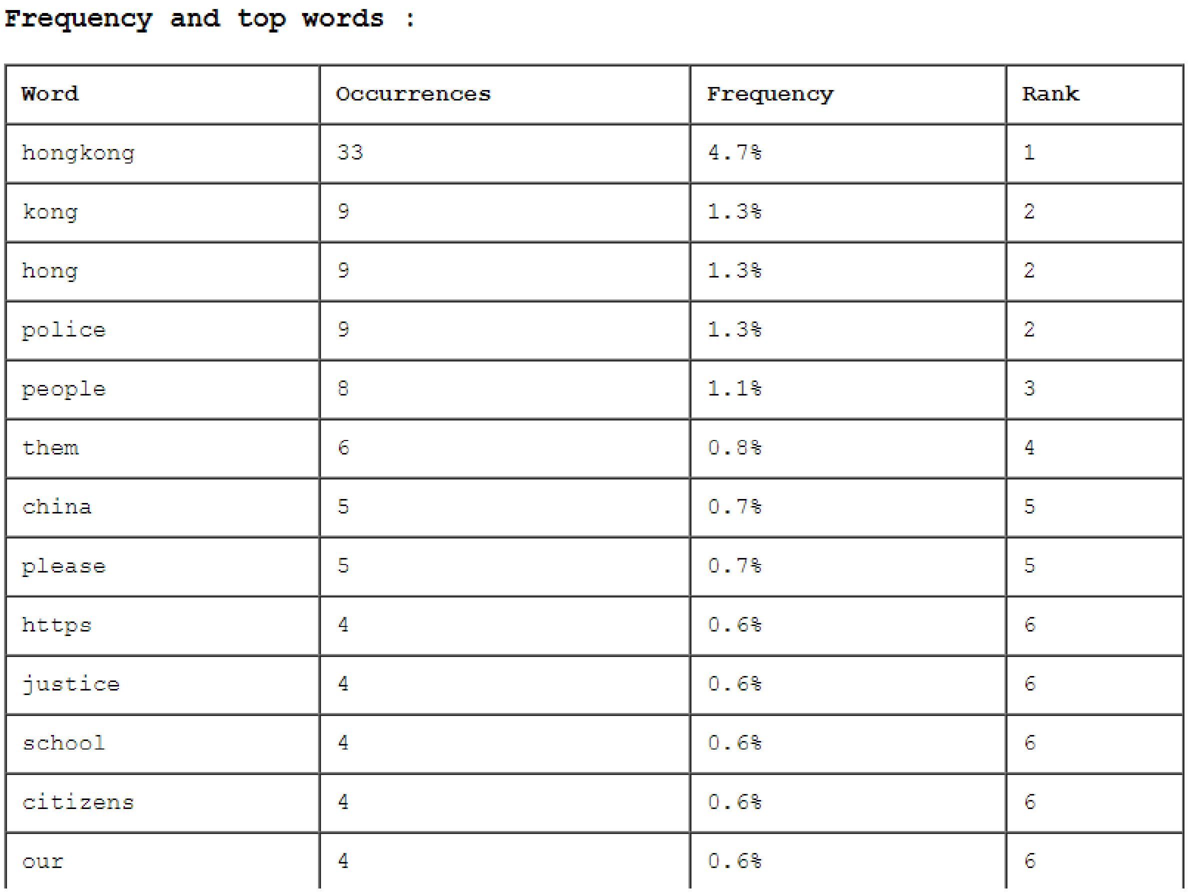
7. china (5, 0.7%);

8. please (5, 0.7%);

9. https (4, 0.6%);

10. justice (4, 0.6%).

The results might also be seen of the figure 3.12:



Frequency and top words :

Word	Occurrences	Frequency	Rank
hongkong	33	4.7%	1
kong	9	1.3%	2
hong	9	1.3%	2
police	9	1.3%	2
people	8	1.1%	3
them	6	0.8%	4
china	5	0.7%	5
please	5	0.7%	5
https	4	0.6%	6
justice	4	0.6%	6
school	4	0.6%	6
citizens	4	0.6%	6
our	4	0.6%	6

Figure 3.12 – The results of keyword analysis on http://textalyser.net/

For some reason the site has detected “https” as a keyword as well.

Website find-keyword.comcounts words and percentage of every word. It gave us the following result (count and percentage in brackets):

1. the (36, 2.63 %);

2. #hongkong (33, 2.41 %);

3. are (11, 0.80 %);

4. this (11, 0.80 %);

5. &(10, 0.73 %);

6. kong (9, 0.66 %);

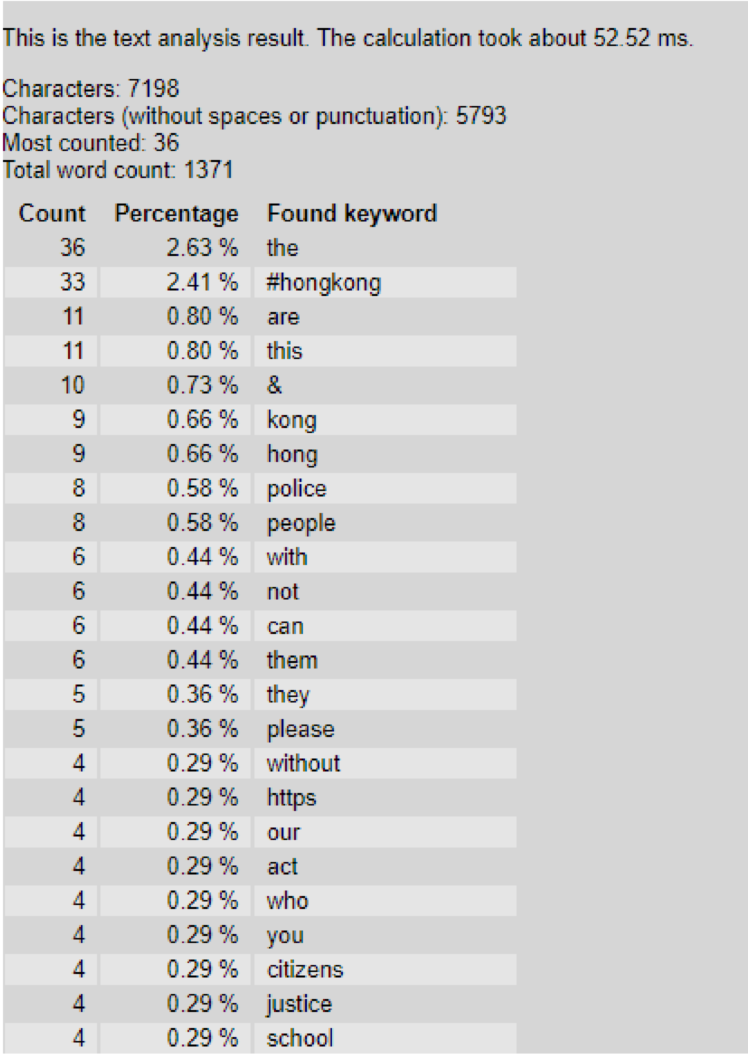
7. hong(9, 0.66 %);

8. police (8, 0.58 %);

9. people (8, 0.58 %);

10. with (6, 0.44 %).

As we may see (figure 3.13), this system has quite poor setting, especially stop words collection. It included results like “the”, “are”, “this” which is inaccurate.



This is the text analysis result. The calculation took about 52.52 ms.

Characters: 7198
Characters (without spaces or punctuation): 5793
Most counted: 36
Total word count: 1371

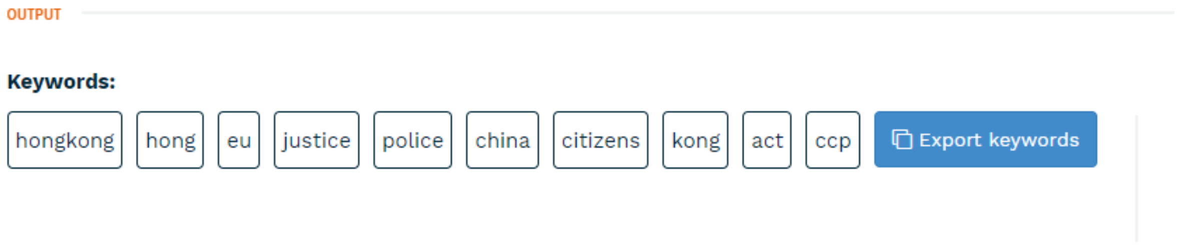
Count	Percentage	Found keyword
36	2.63 %	the
33	2.41 %	#hongkong
11	0.80 %	are
11	0.80 %	this
10	0.73 %	&
9	0.66 %	kong
9	0.66 %	hong
8	0.58 %	police
8	0.58 %	people
6	0.44 %	with
6	0.44 %	not
6	0.44 %	can
6	0.44 %	them
5	0.36 %	they
5	0.36 %	please
4	0.29 %	without
4	0.29 %	https
4	0.29 %	our
4	0.29 %	act
4	0.29 %	who
4	0.29 %	you
4	0.29 %	citizens
4	0.29 %	justice
4	0.29 %	school

Figure 3.13 – The result of analysis on http://textalyser.net/

http://www.cortical.io/ also has a function of keywords extraction. It does not show any metrics but builds a visual finderprint. It shows only top-10 results. The output of check is:

1. hongkong;
2. hong;
3. eu;
4. justice;
5. police;
6. china;
7. citizens;
8. kong;
9. act;
10. ccp.

Words collection seems good, although it is unclear what was meant by words “act” and “ccp”.



Keywords:
hongkong hong eu justice police china citizens kong act ccp

Figure 3.14 – The result of analysis on http://www.cortical.io/

It is pretty obvious that the result we obtained from out application is better than the results of the websites analyzed. Our application includes to the keywords list only nouns and does not add any mess words, so-called “stop words”.

4 MARKETING STRATEGY

Telegram has repeatedly stated that there will be no native advertising in the system, while there are more than 100 thousand bots in it. Already there are ad networks based on the bots of this platform; There are several directories (storebot.me, botsfortelegram.com).

As it was announced by Pavel Durov at TechCrunch Disrupt 2015, in 2016 Payment API to Telegram has been added. It allows bots to accept money from users – both in payment of their services, and in favor of external services. A day before the announcement of the Facebook Messenger Platform for Bots, Telegram rolled out a massive update of its Bots API.

Slack launched an investment fund worth $ 80 million: money will go to third-party teams involved in the development of bots for this platform.

Microsoft announced the Skype Bot SDK and plans to add its bots to all major platforms a couple of weeks before the Facebook conference. Following Microsoft, another major Kik messenger announced the opening of a bot store.

Thus, in this very promising market, there are more and more players, technologies are moving forward faster. The range of tasks that can be solved by bots is much wider than at first glance.

Support services, consultants in online stores, legal and any other advice, mass notification of people if necessary, and much, much more. As a recent example of the Flibusta online library showed, in some cases the bot can even successfully replace the entire portal.

Ross Simmonds, a businessman from Canada, has examined this issue and presented seven profitable business models, allowing to monetize bots.

Model 1: service B2B bots (BaaS).

B2B bots control the task performance process, solve company’s intercommunication issues, help to enter the new efficiency level, and provide many other options.

Ross Simmonds told that revenue from using service applications (SaaS) and networked storages was more than $32 billion previous year. Corporate users have already got used to the similar business model, so the BaaS platform implementation should not cause any problem.

Model 2: bots for native and affiliate advertising.

Currently, a lot of companies pay media resources for posting and distributing their content. But they can use bots for such purposes. For instance, a cooking helper prompts a housewife how to cook certain dishes and recommends to buy Sriracha sauce for better taste, simultaneously showing the reference to the article like “Recipes with Sriracha sauce to impress your family”. Surely, the article will be paid by the company producing this sauce. According to statistics, such native advertising is much more efficient than traditional banners.

Model 3: bots for affiliate marketing.

Affiliate marketing is a perfect earnings strategy for young companies. Chatbots implementation significantly simplifies the process of promoting third party products. In particular, Kip shopping bot responds to user requests regarding where to buy a certain product or order services, and Kip developers obtain the interest for each purchased product from the manufacturing company.

Model 4: bots for conducting paid surveys.

Today, there are bots that can conduct surveys. Thus, owners of such chatbots can use them to provide services of researching tastes and preferences of the target audience. DisOrDatBot is the example of such software.

Model 5: bots for searching potential customers.

Lead generation is a popular marketing method of searching and attracting potential customers. Previously, specialists have created landing pages for such purposes, and nowadays one can implement chatbots that will pay more attention to useful content instead of advertising. For instance, a construction bot will help to choose wallpaper and then transfer the customer information to the company focusing on this area.

Monetization process of botsof this type is the following:

- developers design a special bot for customer searches;
- a bot communicates with customers and collects their data;
- customer data are sent to companies, providing certain services;
- bot developers obtain the interest from product selling.

Model 6: retail bots.

Soon, none of big online shops can do without bots. Surely, it is so convenient: users ask a question whether the website offers a certain product, and a chatbot immediately provide references to required goods. By the way, purchases can be also made via chat. Fast and convenient.

Model 7: expert bot.

People requiring specialist assistance are ready to pay for sound advice or consultation. Modern bots using artificial intelligence algorithms are similar to expert systems, allowing to find the best possible solution of a certain task. It can be a bot answering psychological question, or engineering bot allowing to repair a car on your own, or consulting bot in business management. Chatbots are able to give tips as good as skilled professionals [81].

Any business model will be successful when following three aspects:

1. Efficient and favourable interaction with users.
2. Time and cost saving when using bots.
3. Business and life improvement when following tips from bots.

When choosing a bot promotion strategy, the first thing to do is to focus on its purpose and the audience interested in it. There are several typical groups of chatbots from the point of view of a marketer, this will help determine the choice of promotion strategy (there may not be clear boundaries between them, one robot may have several features):

– bot product. Ready-made application based on chatbot. This includes various services for cropping pictures, photos, search services for friends, games, etc.;

– bot, as the main way to provide services;

– an additional interface of a mobile application, web service, store;

– communication channel with the administration, online consultation;

– an additional marketing channel.

If the development of the robot is planned to create an additional communication channel with the client, it will often be easier and more effective to focus on website promotion, offering to use the Telegram bot on the landing page. The same applies to many highly specialized services, including B2B type (from business to business), large and small projects with an existing distribution network. This path will seem relevant to entrepreneurs for whom a chat robot is not a way to stand out from the crowd, but simply provide a different method of interaction. Since this market is still very young, many projects can get even greater returns from marketing their bot than from promoting a site or software. This is especially relevant for mobile application developers. The costs of creating an MVP based on the Telegram bot (a test version of the product, designed to test the way for large investments in the project) and the subsequent promotion are significantly lower compared to the costs of developing Android and iOS applications. For these reasons, only those companies that see them as the main tool for marketing and providing the final service are engaged in serious promotion of chat bots. The competition among Telegram bots is currently much weaker, which opens up wide doors for start-ups and small businesses, small groups of novice developers.

The cost of attracting new users in the Telegram environment is now lower than the cost of similar services in other areas and, in particular, other messengers. At the same time, Durov’s platform offers better promotion opportunities than its competitors, which makes it stand out in a favorable light. The greatest conversion (return on investment in promotion) is achieved by advertising in other bots, Telegram channels. And this is obvious: a person who does not have a Telegram account is unlikely to follow the link to the bot. For this reason, any content marketing should be highly targeted. Ideally - based on blogs and posts in the social. Telegram networks of subjects.

If the task is to promote an entertaining bot, then "all means are good enough". Excellent results can be achieved from publication in directories. They also work out viral techniques, content marketing, smm publications on social networks. You need to focus on an audience familiar with Telegram. For highly specialized projects designed for a small audience, the best option is a landing page with a link in the context of a chat bot, targeted content marketing. For most projects, advertising works fine in other Telegram bots. It may be convenient for entrepreneurs providing services to the public to distribute information in the form of business cards, product labels, digital signage. To improve advertising policy, it is extremely desirable (even mandatory) to activate the statistics services.

The best option for the bot that we have developed would be to advertise it in one set with the Android application that has been developed, since this kind of bots has specific audience and will not be suitable for ordinary user.

CONCLUSIONS

Natural language processing is a very powerful field of computer science nowadays. It is high technology that is used in every modern smart application – as the problem of "explaining" natural language to machines is the key conception for all devices that we are interacting with on the daily basis. The variety of tasks that it solves is really impressive – from speech recognition to machine translation, from social networks analysis to dialogue systems and many more. NLP is also a very complicated sphere – it requires strong analytical skills, high math and computation proficiency, linguistic and machine learning knowledge. But thanks to modern libraries and tools NLP is now available for people with different background. Social networks are the breeding ground for NLP approaches implementation and testing – they are a platter of information, personalities, social relations, communications, they can be considered a mini-model of the world – they are like mirrors, that immediately reflect everything that happens in the whole planet. In past decades it was common to apply NLP features only to some pieces of art, things, written / created by great minds. But it is in the past nowadays – now the speech or texts of ordinary people are of the same value as analysed they can give a lot of useful insights. So the interest to social media is only growing and for sure will continue growing in the next decades since the number of their users' increases every year. In the work we described different approaches to social media analysis and presented mathematical models for each of them. There were also listed diverse indicators that might be useful during social networks researches.

There are a lot of tools for NLP tasks solving. Every popular programming language has its own libraries, frameworks etc. It is kind of ordinary for modern developer – to carry linguistic analysis. To obtain the expected results we had chosen Python as a programming language and TextBlob as a NLP library. It fully corresponded our goals – we wanted to lead a research, but did not want to spend more time when it is possible to spend less.

We made a research and find out that there are three the most popular social networks in the world. These resources are Facebook, Instagram and Twitter. We analysed the APIs of these giants and picked Twitter as a field for experiments. Twitter API used to be very accessible and functional tool – now it is not like this, because Twitter complicated the process of developer's account obtaining. We already had Twitter developer's, so the choice was very obvious. Companies care about the data privacy more and more, but this generally positive tendency makes developers' lives more difficult.

Chatbots are the mainstream of the last few years. The pioneer platform that somehow inspired this new trend was Telegram – a secure messaging application, created by the former owner of VK. Telegram team created sophisticated chatbot API and started the new wave. Nowadays each popular messenger has an API for chatbot development. We decided to use Telegram as it is one step ahead of the competitors and has not only great API, but is a handy application by itself.

The chatbot that was designed combines a few interesting functions – semantic and sentimental analysis. It also uses a few API and a few libraries. It may be useful for some marketing or social researches, it suggests a semantic analysis that works better that top-3 keyword extract tools offered by Google search engine. Also, an application is just a perfect introduction to the world of chatbots for its developer. We were unable to find analogues to the app so we can assume applications of such profile are not easily available.

REFERENCES

1. СтаднійО. Ю. Автоматичнакласифікаціяучасниківсоціальнихмережнаосновілінгвістичноганалізумікроблогів / О. Ю. Стадній // Матеріалидоповідей XLVII науково-технічноїконференціїпідрозділівВінницькогонаціональноготехнічногоуніверситету, 21–23 березня. – Вінниця :ВНТУ, 2018.

2. Близнюк Б. О. Современные методы обработки естественного языка / Б. О. Близнюк, Л. В. Васильева, И. Д. Стрельников, Д. С. Ткачук // Вісник Харківського національного університету імені В. Н. Каразіна. – 2017. – С. 14-26.

3. Стадній О. Ю. Automated classification of social media users on the basis of microblogs' linguistic analysis / О. Ю. Стадній // Всеукраїнська науково-практична інтернет-конференція «Молодь в науці:дослідження, проблеми, перспективи», Вінниця: ВНТУ, 2019.

4. Manning C. Foundations of Statistical Natural Language Processing / C. Manning, H. Schütze – The MIT Press, 1 edition, 1999 – 620 p.

5. Text categorization: article. URL: http://www.scholarpedia.org/article/Text_categorization.

6. IyyerM., Manjunatha V.,Boyd-Graber J.,DaumeH.Deep Unordered Composition Rivals Syntactic Methods for Text Classification, 2015. URL:https://cs.umd.edu/~miyyer/pubs/2015_acl_dan.pdf

7. Johnson R., Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks, 2014. URL:https://arxiv.org/pdf/1412.1058.

8. BengioY., Ducharme R., Vincent P., Jauvin C.A Neural Probabilistic Language Model. Journal of Machine Learning Research. 2003. P. 1137–1155.

9. Goodfellow I.,BengioY., Courville A.Deep Learning (Adaptive Computation and Machine Learning series): The MIT Press, 2016.800 p.

10. Graves A., Fernandez S., Gomez F., Schmidhuber J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. 2006.URL: http://www.cs.toronto.edu/~graves/icml_2006.pdf.

11. Graves A., Mohamed A., Hinton G. Speech Recognition with Deep Recurrent Neural Networks. 2013. URL: https://arxiv.org/pdf/1303.5778.

12. Abdel-Hamid O., Deng L., Dong Y. Exploring convolutional neural network structures and optimization techniques for speech recognition. 2014.URL: https://www.microsoft.com/en-us/research/publication/exploring-convolutional-neural-network-structures-and-optimization-techniques-for-speech-recognition/.

13. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R., Bengio Y.Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. URL:https://arxiv.org/pdf/1502.03044.

14. Vinyals O., Toshev A., Bengio S., Erhan D. Show and tell: A neural image caption generator. 2015. URL: https://arxiv.org/pdf/1411.4555.

15. Venugopalan S., Rohrbach M., Donahue J., Mooney R., Darrell T., Saenko K. Sequence to Sequence – Video to Text. 2015.URL: https://arxiv.org/pdf/1505.00487.

16. Sutskever I., Oriol V., Le Q. Sequence to Sequence Learning with Neural Networks. 2014. URL: https://arxiv.org/pdf/1409.3215.

17. Bahdanau D., Cho K., Bengio Y Neural Machine Translation by Jointly Learning to Align and Translate, 2014. URL: https://arxiv.org/pdf/1409.0473.

18. Auli M., Galley M. Joint Language and Translation Modeling with Recurrent Neural Networks. 2013. URL:https://www.microsoft.com/en-

us/research/publication/joint-language-and-translation-modeling-with-recurrent-neural-networks/.

19. Rush A., Chopra S. A Neural Attention Model for Abstractive Sentence Summarization. 2015. URL:https://ui.adsabs.harvard.edu/abs/2015arXiv150900685R.

20. Cheng J., LapataM. Neural Summarization by Extracting Sentences and Words.2016. URL:https://arxiv.org/abs/1603.07252

21. NallapatiR., Zhou B. Abstractive Text Summarization Using Sequence-to-Sequence RNNs.2016. URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160206023N.

22. Hermann K.,Kocisky T.,Teaching Machines to Read and Comprehend. 2015URL: http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf.

23. Zong C., Strube M.Question Answering over Freebase with Multi-Column Convolutional Neural Networks. 2015.URL: https://www.aclweb.org/anthology/P15-1026.pdf.

24. Yu L., Hermann K., Blunsom P., PulmanS. Deep Learning for Answer Sentence Selection. 2015.URL: https://arxiv.org/pdf/1412.1632

25. 8 best Python Natural Language Processing (NLP) libraries. URL:https://sunscrapers.com/blog/8-best-python-natural-language-processing-nlp-libraries/.

26. Zelle J. Python Programming: An Introduction to Computer Science (3rd Edition) / J. Zelle. – Franklin, Beedle& Associates, 2010 – 432 p.

27.Natural Language Processing Tools. URL: https://opensource.com/article/19/3/natural-language-processing-tools

28. 8 Best Python NLP Libraries. URL: https://sunscrapers.com/blog/8-best-python-natural-language-processing-nlp-libraries/.

29. Natural Language Toolkit. URL:http://www.nltk.org/. ¶

30. Industrial-Strength Natural Language Processing. URL:https://spacy.io/.

31. TextBlob: Simplified Text Processing. URL:https://textblob.readthedocs.io/en/dev/.

32. textacy 0.9.1. URL: https://pypi.org/project/textacy/

33. Welcome to Pytorch-NLP's documentation! URL: https://pytorchnlp.readthedocs.io/en/latest/

34. RaRe-Technologies/genism. URL:https://github.com/RaRe-Technologies/gensim

35. Welcome to polyglot's documentation! URL: https://polyglot.readthedocs.io/en/latest/index.html.

36. scikit-learn Machine Learning in Python. URL:https://scikit-learn.org/.

37. Pattern. URL:https://www.clips.uantwerpen.be/pages/pattern.

38. OrsiniL.What You Need To Know About Node.js. 2013. URL: http://readwrite.com/2013/11/07/what-you-need-to-know-about-nodejs.

39. Node.js Foundation. URL: https://nodejs.org/foundation/.

40. Linux Foundation Collaborative Projects. URL: http://collabprojects.linuxfoundation.org/.

41. retext. URL:https://www.npmjs.com/package/retext.

42. Compromise. URL:https://www.npmjs.com/package/compromise.

43. Natural. URL:https://www.npmjs.com/package/natural.

44. node-nlp. URL:https://www.npmjs.com/package/node-nlp.

45. Write once, run anywhere? URL: http://www.computerweekly.com/Articles/2002/05/02/186793/write-once-run-anywhere.htm.

46. Design Goals of the Java™ Programming Language. URL: https://www.oracle.com/technetwork/java/intro-141325.html

47. The 10 most popular programming languages, according to the 'Facebook for programmers.*Business Insider*.URL:https://www.businessinsider.de/the-10-most-popular-programming-languages-according-to-github-2018-10?op=1.

48. JavaOne 2013 Review: Java Takes on the Internet of Things. 2016,. URL:https://www.oracle.com/technetwork/articles/java/afterglow2013-2030343.html.

49. Why should I uninstall older versions of Java from my system? URL: https://www.java.com/en/download/faq/remove_olderversions.xml.

50. Welcome to Apache OpenNLP. URL:https://opennlp.apache.org/.

51. Stanford CoreNLP – Natural language software. URL:https://stanfordnlp.github.io/CoreNLP/.

52. CogComp/cogcomp-nlp. URL:https://github.com/CogComp/cogcomp-nlp.

53. NLP framework for JVM languages. URL:https://github.com/emorynlp/nlp4j.

54. Apache UIMA. URL: http://uima.apache.org/.

55. GATE Embedded. URL:https://gate.ac.uk/family/embedded.html.

56. LingPipe. URL:http://alias-i.com/lingpipe/.

57. Social networking service.URL: https://en.wikipedia.org/wiki/Social_networking_service.

58. Most famous social network sites worldwide as of September 2017, ranked by number of active users (in millions). URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

59. Facebook dominates social landscape. URL: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/attachment/12d969c8-2608-4bb0-bb7d-75b796848d0f-1/.

60. Do YOU know what social media rules you've signed up to? URL: https://www.bbc.co.uk/newsround/41426106.

61. Instagram Stories is Now Being Used by 500 Million People Daily. URL:https://www.socialmediatoday.com/news/instagram-stories-is-now-being-used-by-500-million-people-daily/547270/.

62. Instagram Graph API. URL:https://developers.facebook.com/docs/instagram-api.

63. Does Facebook Really Know How Many Fake Accounts It Has? *The New York Times.* 2019. URL: https://www.nytimes.com/2019/01/30/technology/facebook-fake-accounts.html.

64. HuntE. Disputed by multiple fact-checkers: Facebook rolls out new alert to combat fake news. *The Guardian.*URL: https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news.

65. GraphAPI. URL: https://developers.facebook.com/docs/graph-api.

66. FAQ. URL:https://developers.facebook.com/docs/graph-api/faq.

67. Twitter via SMS FAQ. URL:https://help.twitter.com/en/using-twitter/twitter-sms-faqs .

68. About Twitter. URL:https://about.twitter.com/company .

69. Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City. URL:https://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/ .

70. Top Sites. URL:http://www.alexa.com/topsites.

71. Twitter overcounted active users since 2014, shares surge on profit hopes. URL: https://www.usatoday.com/story/tech/news/2017/10/26/twitter-overcounted-active-users-since-2014-shares-surge/801968001/.

72. For Election Day Influence, Twitter Ruled Social Media. URL:https://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter-ruled-social-media.html .

73. About Twitter's APIs. URL:https://help.twitter.com/en/rules-and-policies/twitter-api.

74. РозробкаЧат-Ботів. URL:https://evergreens.com.ua/ua/articles/bot-reviews.html.

75. Messenger Developers API | Official website | Build a bot. URL:https://developers.facebook.com/products/messenger/.

76. Viber API Documentation 7.3.0. URL: https://developers.viber.com/docs/api/rest-bot-api/.

77. Telegram Bot API – Telegram APIs. URL: https://core.telegram.org/bots/api.

78. API Reference - Bot Service | Microsoft Docs. URL:https://docs.microsoft.com/ru-ru/azure/bot-service/rest-api/bot-framework-rest-connector-api-reference.

79. API для чат-ботов | Разработчикам | Вконтакте.URL:https://vk.com/dev/bots_docs.

80. Enabling interactions with bots. URL:https://api.slack.com/bot-users.

81. 7 бизнесс-моделей для заработка на чатботах. URL:https://chatbotconf.com.ua/en/article/7-biznes-modeley-pozvolyayushchih-zarabotat-na-chat-botah-63593.

ATTACHMENTS

AttachmentA

(mandatory)

VNTU

APPROVED BY

Head of the department of the CCS VNTU,

Dr. of Sci. Prof

___________________R. Kventyi

“__” ___________ 2019

TECHNICAL TASK

for the master's qualification work

«Information system of intellectual analysis ofnatural-language content of social networks. Part 2. Automated Classification of Social Media Users on the Basis of Microblogs’ Linguistic Analysis»

08-02.MQW.009.00.000 TT

Supervisor of Masters qualification work

Professor of

the department of the AIITO. Bisikalo

“___” _______________

2019

Performer: st. of the group 3ACIT-18m

O. Stadnii

“___” _______________ 2019

Vinnytsia 2019

1. Name and field of application

   1.1. Title – «Information system of intellectual analysis ofnatural-language content of social networks. Part 2. Automated Classification of Social Media Users on the Basis of Microblogs' Linguistic Analysis»

   1.2. Field of application – Chatbot application for social networks analysis.

2. The basis for the development.

   Theme of master's qualification was approved by order VNTU №____ from "__________"

3. Purpose and designation.

   The purpose of master's thesis is to create the chatbot that would analyze users' profiles, news and reactions of people in social networks to the important events.

4. Initial data for development.

   Master's qualification work is carried out for the first time. During the development, the following documents should be used:

   1. Zelle J. Python Programming: An Introduction to Computer Science (3rd Edition) / J.Zelle. – Franklin, Beedle& Associates, 2010 – 432 p.
   2. Manning C. Foundations of Statistical Natural Language Processing/ C. Manning, H.Schütze – The MIT Press, 1999 – 620 p.

5. Requirements for development.

   5.1. List of main functions:

   - Visualization of information inside of Telegram bot.
   - Function that takes data from Twitter.
   - Sentiment analysis and natural language processing of users' tweets.
   - Possibility of choosing the type of analysis.

   5.2. Basic technical requirements for development.

   5.2.1. Requirements for the software platform: no.

   5.2.2. Operating system conditions:

   - work on mobile devices based on Android 4.0/iOS 9.0 or higher;

   -work on computers.

6. Stages of development.

6.1 Explanatory note:

1	Analysis of the NLP tasks and social network models	
2	Analysis of the problems of social network analysis and NLP, reviewing of appropriate tools	

3	Practical implementation and analysis of the results	
4	Approval of research results	
5	Publications	
6	Designing an explanatory note, graphic material and presentation	
7	Master’s work defense	

7. The order of control and acceptance.

7.1. The progress of master's qualification work is supervised by the head of the work. The border control should be conducted until _________________.

7.2. The project certification is carried out on preliminary protection. Preliminary defense of master's qualification work is to be held until _____________.

7.3. The final decision on the assessment of the quality of the master's qualification work is taken at a meeting of the DEC. The defense of master's qualification work is to hold _______________.

Attachment B (mandatory). List of graphical materials

Head of the department of the AIIT ___________ Dr. of Sci. Prof R. Kvetnyi

(signature) (*name, surname, degree, academic status*)

Scientific supervisor ___________ Dr. of Sci. Prof O. Bisikalo

(signature) (*name, surname, degree, academic status*)

Technical control ___________ Dr. of Sci. Prof O. Bisikalo

(signature) (*name, surname, degree, academic status*)

Regulatory control ___________ Dr. of Sci. Prof O. Bisikalo

(signature) (*name, surname, degree, academic status*)

Reviewer ___________ _________________ _________________

(signature) (*name, surname, degree, academic status*)

Student of the 3ACIT-18m ___________ O. Stadnii

(signature) (*name,surname*)

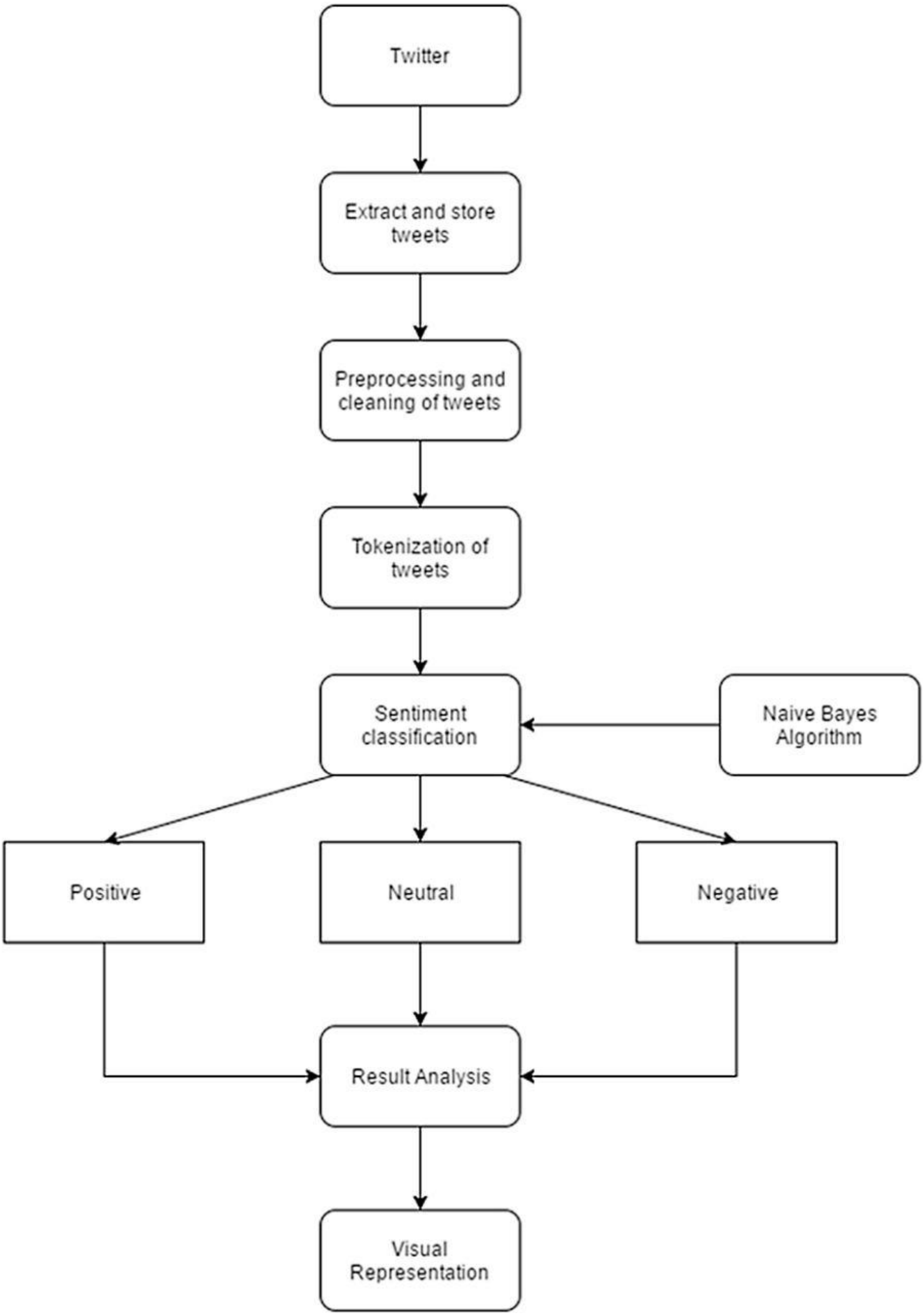


Figure B.1 – Sentiment analysis algorithm

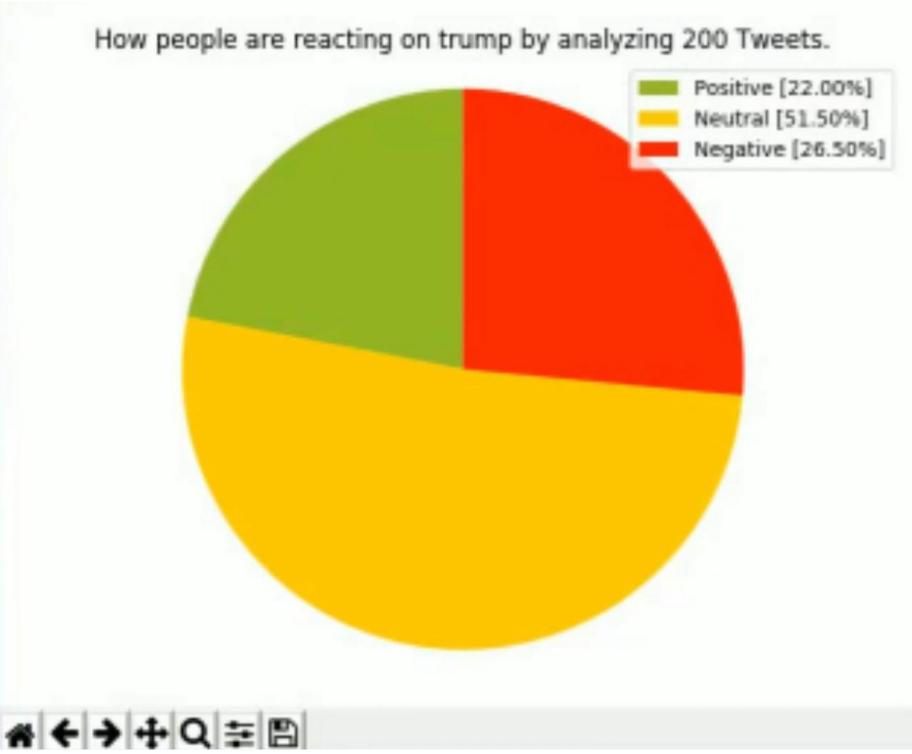


Figure B.2 – Sentiment analysis diagram

Attachment C (mandatory). Listingof the program

Module of sentiment analysis.

```
import sys
import csv
# import pygraphviz as pgv
import numpy as np
import networkx as nx
from datetime import datetime
import math
from collections import defaultdict
importliwc.categories

# importliwc.categories

labels = ["first_person", "second_person", "third_person", "posemo", "negemo", "cognitive",
"sensory", "time", "past",
        "present", "future", "work", "leisure", "swear", "social", "family", "friend", "humans",
"anx", "anger",
        "sad", "body", "health", "sexual", "space", "time", "achieve", "home", "money", "relig",
"Affect", "cause",
        "Quant", "Numb", "inhib", "ingest", "motion", "nonfl", "filler",
"number_classified_words", "number_words"]



def main():
tweets_csv = sys.argv[1]
output_file = open('liwc_' + tweets_csv, 'w')
csv_writer = csv.writer(output_file, delimiter=',')

   with open(tweets_csv) as csvDataFile:
csvReader = csv.reader(csvDataFile)
```

```
        for row in csvReader:
tweet_id = row[0]
tweet_date = row[1]
tweet_text = row[2]

            print(tweet_text)
liwc_data = liwc_classify(tweet_text)

            row = [tweet_id, tweet_date, tweet_text] + liwc_data
csv_writer.writerow(row)



defliwc_classify(text):
    # text is the body of text we want to run LIWC on
vals = liwc.categories.classify(text)

    for i in range(len(vals)):
        print(labels[i], vals[i])

    return list(vals)



if __name__ == "__main__":
main()
```

Tweets fetch module.

```
import tweepy #https://github.com/tweepy/tweepy
import csv

#Twitter API credentials
```

```
consumer_key = ""
consumer_secret = ""
access_key = ""
access_secret = ""



defget_all_tweets(screen_name):
        #Twitter only allows access to a users most recent 3240 tweets with this method

        #authorize twitter, initialize tweepy
        auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_key, access_secret)
        api = tweepy.API(auth)

        #initialize a list to hold all the tweepy Tweets
        alltweets = []

        #make initial request for most recent tweets (200 is the maximum allowed count)
        new_tweets = api.user_timeline(screen_name = screen_name,count=200)

        #save most recent tweets
        alltweets.extend(new_tweets)

        #save the id of the oldest tweet less one
        oldest = alltweets[-1].id - 1

        #keep grabbing tweets until there are no tweets left to grab
        while len(new_tweets) > 0:
                print "getting tweets before %s" % (oldest)
                #all subsiquent requests use the max_idparam to prevent duplicates
                new_tweets = api.user_timeline(screen_name = screen_name,count=200,max_id=oldest)

                #save most recent tweets
                alltweets.extend(new_tweets)

                #update the id of the oldest tweet less one
                oldest = alltweets[-1].id - 1

                print "...%s tweets downloaded so far" % (len(alltweets))

        #transform the tweepy tweets into a 2D array that will populate the csv
        outtweets = [[tweet.id_str, tweet.created_at, tweet.text.encode("utf-8")] for tweet in alltweets]

        #write the csv
        with open('%s_tweets.csv' % screen_name, 'wb') as f:
                writer = csv.writer(f)
                writer.writerow(["id","created_at","text"])
                writer.writerows(outtweets)

        pass
if __name__ == '__main__':
        #pass in the username of the account you want to download
        get_all_tweets("Pontifex")
```