Semantic-Based Linguistic Platform for Big Data Processing

Bobkov A. *) , Gafurov S. **), Krasnoproshin V.*), Vissia H.**)

*)Belarusian State University, 4 Nezavisimosti av., Minsk, 220030, Republic of Belarus, e-mails: anatoly.bobkov@gmail.com , krasnoproshin@bsu.by
**) ByeleX BV Argon 1 – 4751 XC Oud Gastel, The Netherlands, e-mails: gafurov@gmail.com, *h.vissia@byelex.com*

Abstract: *The paper deals with the development of a semantic-based linguistic platform. Special attention is paid to semantic patterns.*

Keywords: big data, natural language processing, semantic patterns, ontology-based approach.

1. Introduction

Big data has been a widely discussed topic for the past five years [1, 2, 3, 4, 5]. The term "Big Data" refers to the large amounts of data in which traditional data processing procedures and tools would not be able to handle. The idea is that mass quantities of gathered data give us unprecedented insights and opportunities across all industries and businesses for solving problems and decision making. Big data is not only an area of potential innovation but is also a crucial factor that companies address to survive in the modern marketplace.
There's no doubt that big data will continue to play an important role in many different industries around the world.
Currently, information extraction from big data becomes predominant. The information can come from various sources, e.g. media, blogs, personal experiences, books, newspaper and magazine articles, expert opinions, encyclopedias, web pages, etc.
Today, big data gives us unprecedented insights and opportunities across all industries from healthcare to financial to manufacturing and more.
Businesses can make a lot out of big data, making it an important resource.
The use and adoption of big data within governmental processes allows efficiencies in terms of cost, productivity, and innovation.

2. Problem statement and solution

Information extraction from big data comprises methods, algorithms and techniques for finding the desired, relevant information and for storing it in appropriate form for future use.
The field of information extraction is well suited to various types of business, government and social applications [6, 7]. Diverse information is of great importance for decision making on products, services, events, persons, organizations.
Creation of systems that can effectively extract meaningful information requires overcoming a number of challenges: identification of documents, knowledge domains, specific opinions, opinion holders, events, activities, mood state, as well as representation of the obtained results.
Numerous models and algorithms are proposed for web information processing and information extraction [8, 9]. But traditional data processing technologies and tools are not able to adequately deal with large amounts of data. Big data is too voluminous and requires the use of new technologies and data-processing applications to effectively capture, store, analyze, and present big data. Thus, the problem of effective information extraction from texts in a natural language still remains unsolved.

The purpose of this paper is to describe the developed and integrated semantic-based linguistic platform for solving the problem of effective extraction of meaningful, user-oriented information from big data.

Semantic relations [10, 11] play the major role in extracting meaningful information. Semantic relations (lexical-semantic relations) are meaningful associations between two or more concepts or entities. They can be viewed as links between the concepts or entities that participate in the relation. Associations between concepts can be categorized into different types.

In information extraction and text mining, word collocations show a great potential [12] to be useful in many applications (machine translation, natural language processing, lexicography, word sense disambiguation, etc.).

"Collocations" are usually described as "sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent" [13].

The traditional method of performing automatic collocation extraction is to find a formula based on the statistical quantities of words to calculate a score associated to each word pair. The formulas are mainly: "mutual information", "t-test", "z test", "chi-squared test" and "likelihood ratio" [14].

Word collocations from the point of semantic constituents have not yet been widely studied and used for extracting meaningful information, especially when processing texts in a natural language.

The developed approach is based on word collocations on the semantic level and contextual relations forming a "semantic pattern".

A semantic pattern is a kind of a knowledge model. Knowledge modeling describes what data means and where it fits. It allows us to understand how different pieces of information relate to each other. A semantic pattern can be viewed as containing slots that need to be filled. Though most patterns are binary ones having two slots, a pattern may have three or more slots.

Semantic patterns help users to ask questions in a natural way and discover relationships between disparate pieces of information.

In general, the proposed semantic patterns include: 1) *participants* (a person, company, natural/manufactured object, as well as a more abstract entity, such as a plan, policy, etc.) involved in the action or being evaluated; 2) *actions* - a set of verb semantic groups and verbal nouns ("buy", "manufacture", "arrival", etc.); 3) *rules for semantic patterns actualization*.

The patterns cover different types of semantic relations: 1) semantic relations between two concepts/entities, one of which expresses the performance of an operation or process affecting the other ("Much remains to be learned about how nanoparticles affect the environment"); 2) synonymous relationships ("beautiful – attractive - pretty"); 3) antonymy ("wet - dry"); 4) causal relations ("Research identifies new gene that causes osteoporosis"); 5) hyponymous relations ("Jaguar is a powerful vehicle"); 6) locative relations ("Amsterdam is located in the western Netherlands, in the province of North Holland"); 7) part-whole relations ("car transmission - car"); 8) semantic relations in which a concept indicates a time or period of an event designated by another concept ("Second World War, 1939-1945"); 9) associative relations ("baker – bread": "The baker produced bread of excellent quality"); 10) "made-of" relations ("This ring is made of gold"); 11) "made-from" relations ("Cheese made from raw milk imparts different flavors and texture characteristics to the finished cheese"); 12) "used-for" relations ("Database software is used for the management and storage of data and databases"); 13) homonym relations ("bank of the river – bank as a financial institution"), etc. A semantic relation can be expressed in many syntactic forms. Besides words, semantic relations can occur at higher levels of text (between phrases, clauses, sentences and larger text segments), as well as between documents and sets of documents. The variety of semantic relations and their properties play an important role in web information processing for extracting relevant fragments of information from unstructured text documents.

An ontology-based approach is used for semantic patterns recognition and extraction [15].

Ontologies have become common on the World-Wide Web [16]. The broadened interest in ontologies is based on the feature that they provide a machine-processable semantics of information sources that can be communicated among agents as well as between software artifacts and humans. More recently, the notion of ontologies has attracted attention from such fields as intelligent information integration, cooperative information systems, information retrieval, electronic commerce, and knowledge management. For any given knowledge domain, an ontology represents the concepts which are held in common by the participants in a particular domain. Since ontologies explicitly represent knowledge domain semantics (terms in the domain and relations among them), they can be effectively used in solving information extraction problems.

Semantic patterns and the ontology-based approach form the basis for the developed semantic-based linguistic platform (Fig. 1). The platform is a group of technologies that are used as a base upon which applications, processes and technologies are developed.

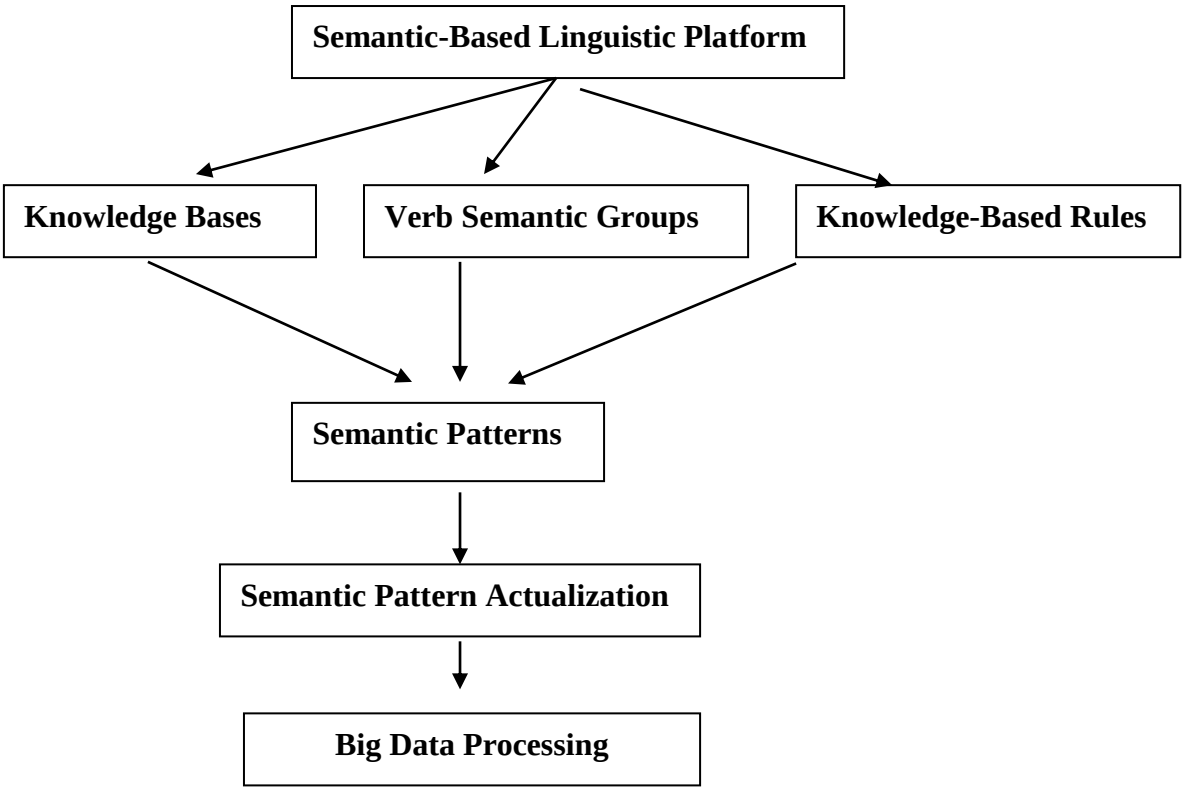

Fig. 1. Semantic-based linguistic platform

3. Implementation of the linguistic platform

The developed semantic-based linguistic platform has been successfully realized in BuzzTalk portal [17]. BuzzTalk gathers meaningful information from extensive sources of textual information (i.e. news, scientific articles, Web pages, tweets, reports, online encyclopedias, etc.) on the level of single words, phrases and sentences. BuzzTalk is offered to companies as a SaaS (Software as a Service) model.

The difference between a traditional search engine and a discovery engine such as BuzzTalk, is that search engines list all results for a specific search whereas BuzzTalk:

- allows you to monitor topic-specific developments within your search;
- discovers the latest information about a particular brand, competitors or industry, thus facilitating to make better decisions;
- collects all text documents from over 58 000 of the most active websites around the globe, two thirds are news sites and one third is blog sites. The authors of these documents are mainly scientists, journalists and opinion leaders;
- finds and links relevant information in natural-language documents while ignoring extraneous, irrelevant information;
- presents a list of articles in chronological order based on publication date. This list grows each day. You can sort and filter this list based on a variety of criteria such as sentiment, mood state, happenings, etc., thus to experience the wealth of real time information without the pain of information overload. For example, you can easily find all publications within your theme that relate to product releases, employment changes, merger & acquisitions and many more.

Below are examples of information extraction in BuzzTalk.

3.1 Economic activities detection

Semantic patterns approach helps to extract information dealing with economic activities. The information could be valuable in many subject areas, including medicine, biology, science, technology, etc.

Recognition of economic activities is closely connected with big data.

The recognition of economic phenomena is rather difficult within natural language processing. Certain elements need to be chosen and grouped according to particular characteristics. Thus, all economic phenomena that are to be described and processed require systematic classification especially when processing big data.

BuzzTalk detects 233 economic activities from texts in a natural language. The economic activities cover all major activities represented in NACE classification (Statistical Classification of Economic Activities in the European Community), which is similar to the International Standard Industrial Classification of all economic activities (ISIC) reflecting the current structure of the world economy. NACE classification provides the internationally accepted standard for categorizing units within an economy. Categories of the classification have become an accepted way of subdividing the overall economy into useful coherent industries that are widely recognized and used in economic analysis, and as such they have become accepted groupings for data used as indicators of economic activities. The classification is widely used, both nationally and internationally, in classifying economic activity data in the fields of population, production, employment, gross domestic product and other. It’s a basic tool for studying economic phenomena, fostering international comparability of data and for promoting the development of sound national statistical systems. The classification provides a comprehensive framework within which economic data can be collected and reported in a format that is designed for purposes of economic analysis, decision-taking and policy-making.

While extracting and analyzing economic activities, BuzzTalk ensures a continuing flow of information that is indispensable for the monitoring, analysis and evaluation of the performance of an economy over time. Moreover, BuzzTalk facilitates information extraction, presentation and analysis at detailed levels of the economy in an internationally comparable, standardized way.

Examples of economic activities detection:

- Toyota has maintained its position as the world's biggest car manufacturer.

Extracted instances:

Economic activities = **Manufacture of motor vehicles** (NACE code C291)

- The world's first auto show was held in England in 1895.

Extracted instances:

Economic activities = **Organisation of conventions and trade shows** (NACE code N823)

- Goat cheese has been made for thousands of years, and was probably one of the earliest made dairy products.

Extracted instances:

Economic activities = **Manufacture of dairy products** (NACE code C105)

- This invention relates to a process for the hardening of metals.

Extracted instances:

Economic activities = **Treatment and coating of metals** (NACE code C256)

- India is the largest grower of rice.

Extracted instances:

Economic activities = **Growing of rice** (NACE code A0112)

- OCBC Bank operates its commercial banking business in 15 countries.

Extracted instances:

Economic activities = **Monetary intermediation** (NACE code K641)

- It is even more important to properly plan the preparation of legal documents.

Extracted instances:

Economic activities = **Legal activities** (NACE code M691)

- Doran Polygraph Services specializes in professional certified polygraph testing utilizing the latest equipment and most current software with techniques approved by the American Polygraph Association.

Extracted instances:

Economic activities = **Security and investigation activities** (NACE code N80)

- Florida's aquafarmers grow products for food (fish and shellfish).

Extracted instances:

Economic activities = **Aquaculture** (NACE code A032)

3.2 Subject domains recognition

In BuzzTalk a subject domain is recognized on the basis of a particular set of noun and verb phrases unambiguously describing the domain. For solving the problem of disambiguation special filters, based on the contextual environment (on the level of phrases and the whole text), are introduced. Subject domains and their concepts are organized hierarchically to state "part-of", "is a kind of" relations.

Examples:

- The Forest Inn Hotel offers hotel accommodation on a weekly basis.

Extracted instances:

Subject domain = **Travel-Hotel**

- The goal of the pollution prevention and reduction program is to prevent or minimize polluting discharges.

Extracted instances:

Subject domain = **Ecology**

- Mozzarella cheese is a sliceable curd cheese originating in Italy.

Extracted instances:

Subject domain = **Food**

- Fresh milk is the common type of milk available in the supermarket.

Extracted instances:

Subject domain = **Beverage**

- Distance education includes a range of programs, from elementary and high school to graduate studies.

Extracted instances:

Subject domain = **Education**

- The biathlon is a winter sport that combines cross-country skiing and rifle shooting.

Extracted instances:

Subject domain = **Biathlon**

- The aim of nanoelectronics is to process, transmit and store information by taking advantage of properties of matter that are distinctly different from macroscopic properties.

Extracted instances:

Subject domain = **Sustainable Business**

- Britain has made a political decision that will have economic effects.

Extracted instances:

Subject domain = **Politics**

- Economy from then on meant national economy as a topic for the economic activities of the citizens of a state.

Extracted instances:

Subject domain = **Economics**

- The law-making power of the state is the governing power of the state.

Extracted instances:

Subject domain = **Law**

- The president called for collective efforts to fight world terrorism.

Extracted instances:

Subject domain = **Terrorism**

- Japan was hit by a magnitude 6.5 earthquake followed by an M7.3 quake on Saturday.

Extracted instances:

Subject domain = **Disaster**

For solving the problem of disambiguation special filters, based on the contextual environment (on the level of phrases and the whole text), are introduced. Subject domains and their concepts are organized hierarchically to state "part-of", "is a kind of" relations.

3.3 Named entities recognition

Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities in a text into pre-defined categories such as names of persons, organizations, locations, etc.

BuzzTalk recognizes the following main named entities:

- "Person" (first, middle, last names and nicknames, e.g. Steve Jobs, Cristina Fernandez de Kirchner);
- "Title" (social, academic titles, etc.);
- "Position" (a post of employment/office/job, e.g. president, CEO);
- "Organization" (a company, governmental, military or other organizations, e.g. Microsoft, Wells Fargo, The University of Oxford);
- "Location" (names of continents, countries, states, provinces, regions, cities, towns, e.g. Africa, The Netherlands, Amsterdam);
- "Technology" (technology names or a description of the technology, e.g. 4D printing, advanced driver assistance, affinity chromatography, agricultural robot, airless tire technology);
- "Product" (e.g. Sikorsky CH-148 Cyclone, Lockheed Martin F-35 Lightning II, Kalashnikov AKS, Windhoek Lager, Mercedes S550, Apple iPhone 6S Plus, Ultimate Player Edition, Adenosine);
- "Event" (a planned public/social/business occasion, e.g. Olympic Summer Games, World Swimming Championship, Paris Air Show, International Book Fair);

- "Industry Term" (a term related to a particular industry, e.g. advertising, finance, aviation, automotive, education, film, food, footwear, railway industries);
- "Medical treatment" (terms related to the action or manner of treating a patient medically or surgically, e.g. vitamin therapy, vaccination, treatment of cancer, vascular surgery, open heart surgery), etc.

The named entities are hierarchically structured, thus ensuring high precision and recall.
For example:
"Organization"

- Airline company
- Automaker
- Bank
- Football club
- Computer manufacturer
- Educational institution
- Food manufacturer
- Apparel manufacturer
- Beverage manufacturer …

3.4 Event extraction

A specific type of knowledge that can be extracted from texts is an event, which can be represented as a complex combination of relations. Event extraction is beneficial for accurate breaking news analysis, risk analysis, monitoring systems, decision making support systems, etc. BuzzTalk performs real-time extraction of 35 events, based on lexical-semantic patterns, for decision making in different spheres of business, legal and social activities. The events include: "Environmental Issues", "Natural Disaster", "Health Issues", "Energy Issues", "Merger & Acquisition", "Company Reorganization", "Competitive Product/Company", "Money Market", "Product Release", "Bankruptcy", "Bribery & Corruption", "Fraud & Forgery", "Treason", "Hijacking", "Illegal Business", "Sex Abuse", "Conflict", "Conflict Resolution", "Social Life", etc.
For example:

- Contract medical research provider, Quintiles, agreed to merge with healthcare information company, IMS Health to make a giant known as Quintiles IMS in an all-stock deal.

Extracted instances:
Event = **Merger & Acquisition**

- Mazda Motor Corporation unveiled the all-new Mazda CX-5 crossover SUV.

Extracted instances:
Event = **Product Release**

- TCS ranked as top 100 U.S. brand for second consecutive year.

Extracted instances:
Event = **Competitive Product/Company**

- Two Hong Kong men arrested for drug trafficking.

Extracted instances:
Event = **Illegal Business**

- A former President of Guatemala, already in jail, has been accused of taking bribes.

Extracted instances:
Event = **Bribery & Corruption**

- Yet another green-energy giant faces bankruptcy.

Extracted instances:
Event = **Bankruptcy**

- Two Afghans held for attempted rape of woman on Paris train.

Extracted instances:
Event = **Sex Abuse**

- A New York woman faced charges for faking cancer to solicit money from unsuspecting donors and a relative.

Extracted instances:
Event = **Fraud & Forgery**

The extracted events play a crucial role in daily decisions taken by people of different professions and occupation.

3.5 Opinion mining

Creation of systems that can effectively process subjective information requires overcoming a number of new challenges: identification of opinion-oriented documents, knowledge domains, specific opinions, opinion holders, representation of the obtained results.
Opinion mining is gaining much popularity within natural language processing ([18]). Web reviews, blogs and public articles provide the most essential information for opinion mining. This information is of great importance for decision making on products, services, persons, events, organizations.
Opinion words are the main constituents of opinion mining and sentiment analysis.
Numerous models and algorithms are proposed to identify and extract opinion words, positive or negative assessment of the object being evaluated [18, 19, 20]. But the problem of effective identification and extraction of opinion words and phrases from an arbitrary text, irrespective of the knowledge domain, still remains unsolved.
We propose an ontology-based approach [15] that helps to identify and process opinion words expressing:
1) appreciation (e.g. flexible, efficient, stable, reduced, ideal, backward, poor, highest)
2) judgement (e.g. active, decisive, caring, dedicated, intelligent, negligent, evil)

While "judgement" evaluates human behaviors, "appreciation" typically deals with natural objects, manufactured objects, as well as more abstract entities, such as plans and policies. Humans may also be evaluated by means of "appreciation", rather than "judgement", when viewed more as entities than as participants, e.g. *lovely medical staff*.
Opinion words can be expressed by: an adjective (*brilliant, reliable*); a verb (*like, love, hate, blame*); a noun (*garbage, triumph, catastrophe)*; a phrase (*easy to use, simple to use*). Adjectives derive almost all disambiguating information from the nouns they modify, and nouns are best disambiguated by directly adjacent adjectives or nouns.
Information about the force of evaluation (low, high, the highest) and orientation (positive/negative) is also taken into consideration. For example, *safe* (low force, positive orientation), *safer* (high force, positive orientation), *the safest* (the highest force, positive orientation), *unsafe* (low force, negative orientation).
Opinion words go together with their accompanying words, thus forming "opinion collocations" (e.g. *deep depression, deep devotion, warm greetings, discuss calmly, beautifully furnished*). By an "opinion collocation" we understand a combination of an opinion word and accompanying words, which commonly occur together in an opinion-oriented text. The use of opinion collocations is a way to solve the problem of opinion word sense disambiguation (e.g. *well-balanced political leader* and *well-balanced wheel*) and to exclude words that do not relate to opinions (cf. *attractive idea* and *attractive energy*).
We assume that the number of opinion collocations, which can be listed in a knowledge base, is fixed.
The use of opinion collocations within the ontology-based approach opens a possibility to assign names of knowledge domains to them, because opinion collocations are generally domain specific. For example, *helpful medical staff* ("health care"), *helpful hotel reception staff* ("travel-

hotel"), *stable economy* ("economics"), *well-balanced politician* ("politics"). More than one knowledge domain may be assigned to an opinion collocation, e.g. *fast service* ("economics-company", "travel-hotel"). Domain-specific information helps to solve the problem of opinion word sense disambiguation and ensures customized search, i.e. detection of sentences relevant to a given knowledge domain or topic.

Processing of the extracted opinion collocations is carried out in their contextual environment. The developed algorithm checks for the presence of modifiers that can change the force of evaluation and orientation indicated in the knowledge base.

Let's consider the following example: *reliable company*. This opinion collocation has the following information in the knowledge base: "low force" of evaluation and "positive orientation".

e.g. Sunpak is a *reliable company.*

The evaluation force is changed to "higher force" in *more reliable company*.

e.g. MSI a *more reliable company*.

The algorithm changes the force of evaluation to "the highest force" when processing the opinion collocations *very reliable company, the most reliable company, extremely reliable company.*

e.g. Electa Limited Company is a *very reliable company.*

The orientation is changed to the opposite ("negative orientation") in the following examples: *unreliable company, not reliable company* ("low force"), *the most unreliable company* ("the highest force").

e.g. With regards to security, Websense may be *the most unreliable company*.

The opinion collocation *not reliable enough company* has positive orientation, but the "low force" is weakened.

e.g. Obviously because InPlant is *not a reliable enough company.*

"The highest force" of evaluation is weakened in *not a very reliable company* ("positive orientation").

e.g. MedZilla is *not a very reliable company* in terms of loyalty to the sales force.

There is also additional information about quality characteristics and relationships for different objects on which an opinion is expressed (e.g. *software product* evaluation includes: usability, reliability, efficiency, reusability, maintainability, portability, testability; *travel-hotel* evaluation includes: value, rooms, location, cleanliness, check in/front desk, service).

Associative relationships, which relate concepts across the tree structure, are also taken into consideration: 1) nominative relationships describing the names of concepts; 2) locative relationships describing the location of one concept with respect to another; 3) associative relationships that represent, for example, the functions, processes a concept has or is involved in; 4) cause-effect relationships.

Based on the proposed ontology approach, an object of the particular class of interest may have its own specific sets of sub-classes, opinion collocations and evaluation. In the automobile domain, for a car model they can be: engine, transmission, suspension, size, color, design,

condition under which an evaluation applies (e.g. driving on slippery roads), a supporting factor for the evaluation.
For example:
"The *C180K* is the *cheapest* in the C-Class range. Everything about the C180K's interior reeks of *superior design* and *craftsmanship*. The *engine revs* as *smoothly* as it sounds. The *steering* was *light*. Whether surmounting cobblestones, concrete or brick, the C180K was a *planted*, *communicative* and *comfortable city car*. The C180K delivers *excellent fuel economy*."
There is also additional information about quality characteristics and relationships for different objects on which an opinion is expressed (e.g. *software product* evaluation includes: usability, reliability, efficiency, reusability, maintainability, portability, testability; *travel-hotel* evaluation includes: value, rooms, location, cleanliness, check in/front desk, service).
For example:
"The *location* of the Golden Well hotel is *excellent*. The *hotel* is *beautifully furnished* without being overdone. *Check-in* was *fast* and *easy*. The *room* was *fabulous*, and the *breakfasts amazing*. The *bed* was *comfortable* and the *bathroom* was a *pleasure*. *Friendly* and *attentive staff*."

The results of opinion collocations processing are grouped and evaluated to recognize the quality of the opinion-related text. The results are also visualized.

3.6 Mood state detection

A valuable addition to opinion mining is detection of individual/public mood states. BuzzTalk mood detection uses the classification of the widely-accepted "Profile of Mood States" (POMS), originally developed by McNair, Lorr and Droppleman. The relationship between mood states and different human activities has proven a popular area of research ([21]).

BuzzTalk mood detection uses the classification of the widely-accepted "Profile of Mood States" (POMS), originally developed by McNair, Lorr and Droppleman ([22]).

In BuzzTalk, mood state detection is based on: 1) mood indicators (e.g. "I feel", "makes me feel", etc.); 2) mood words (e.g. anger, fury, horrified, tired, taken aback, depressed, optimistic); 3) special contextual rules to avoid ambiguation. BuzzTalk automatically recognizes the following mood states: "Anger", "Tension", "Fatigue", "Confusion", "Depression", "Vigor".

For example:

- Despite these problems, I feel very happy.

Extracted instances:
Mood state = **Vigor**

- I'm feeling angry at the world now.

Extracted instances:
Mood state = **Anger**

- I feel fatigued and exhausted.

Extracted instances:
Mood state = **Fatigue**

- I have suicidal thoughts everyday.

Extracted instances:
Mood state = **Depression**

Mood state detection alongside with opinion mining can give answers to where we are now and where will be in future.

3.7 Predictive Analytics

With the use of information technologies a decision maker has a great possibility to know and investigate what is happening, when and where, closely monitor the existing current situation in the world and make predictions. Predictive analytics [23, 24] is used to make predictions about unknown future events. Predictive analytics is used in marketing, financial services, insurance, telecommunications, retail, travel, mobility, healthcare, child protection, pharmaceuticals, capacity planning and other fields. The goal is to go beyond the knowledge of what has happened to provide the best assessment of what will happen in future.

Organizations are turning to predictive analytics to solve difficult problems and uncover new opportunities. Analytical methods can improve crime detection and prevent criminal behavior. As cybersecurity becomes a growing concern, high-performance behavioral analytics examines all actions on a network in real time to spot abnormalities that may indicate fraud, zero-day vulnerabilities and advanced persistent threats. In addition to detection of claims fraud, the health insurance industry is taking steps to identify patients most at risk of chronic disease and to find what interventions are best. Predictive analytics is used to determine customer responses, as well as to promote cross-sell opportunities. Predictive models help businesses to attract, retain and grow their most profitable customers. Many companies use predictive models to forecast equipment failures and future resource needs. Airlines use predictive analytics to set ticket prices. Hotels try to predict the number of guests for any given time to maximize occupancy and increase revenue. Predictive analytics enables organizations to function more efficiently. It helps to detect earthquakes, floods, hurricanes, as well as to forecast future occurrences of such hazards and their various characteristics (magnitude of an earthquake, track and intensity of a cyclone, etc.). Predictive analytics can fix small problems before they become big ones.

Endless flood of Internet data calls for substantial analytical work. Thus, of great importance is the development of effective computer systems for predictive analytics within natural language processing. Predictive analytics, as an area of big data mining, involves extraction of information and its use to predict events, trends, behavior patterns, etc.

We consider that extraction and processing of “cause-effect” relations from texts form the basis for predictive analytics. Knowledge of “cause” and “effect” ensures rational decision making and problem solving. It is important in all areas of science and technology.

A “cause-effect” [25, 26] is a relation in which one event (“cause”) makes another event happen (“effect”). “Effect” is defined as what happened. “Cause” is defined as why something happened.

For example:

- Insecticide lindane found to cause cancer
- Scientists identified signatures of cancer caused by radiation
- Culprit identified as a major cause of vision loss
- Newly discovered molecule could lead to effective treatment for heart failure
- Exposure to phthalates could be linked to pregnancy loss
- Stiff and oxygen-deprived tumors promote spread of cancer
- The car accident was due to the adverse driver's negligence
- The car didn't brake in time. The reason was that the road was slippery

The “cause-effect” relation affects all aspects of our lives. For every “effect” there is a definite “cause”, likewise for every “cause”, there is a definite “effect”. This means that everything that we currently have in our lives is an “effect” that is a result of a specific “cause”.

Though many of the cause-effect relations in texts are implicit and have to be inferred by the reader, the English language actually possesses a wide range of linguistic expressions for explicitly indicating “cause” and “effect” [27, 28, 29].

The following main means are identified:

1) causative verbs (*cause, result in, lead to, make happen, provoke, encourage,* etc.)

For example:

• Lung cancer, brain disease *caused* death
• The crash *resulted in* the deaths of 15 passengers
• E-cigarettes may *lead to* cancer and heart disease

2) causal links (*so, hence, therefore, because of, on account of, that's why, due to, as a result of, owing to, thanks to, by reason of, by cause of,* etc.)

For example:
• Schools are closed *because of* flu
• He knew he could not win the election - *hence* his decision to withdraw
• US Postal Service suspends services *due to* Hurricane Irma

3) conditionals (i.e. "*if …, then*" constructions)

For example:
• *If* the demand for a product is elastic, *then* a business owner can cut the price
• *If* you use correct punctuation, *then* you will include commas where necessary
• *If* an economy is producing efficiently, *then* it is possible for that economy to produce more of one good without producing less of the other.

4) causative nouns (*cause of, reason for, result of, consequence of, influence of, impact of,* etc.)

For example:
• Probably the most serious and most short-sighted *consequence of* deforestation is the loss of biodiversity
• Warm, wet winters during recent decades in the Northern Hemisphere can be explained by the *influence of* greenhouse gases on atmospheric winds
• In coming decades, global warming will have a dramatic *impact on* regional water supplies.
• The most common *cause of* dehydration in young children is severe diarrhea and vomiting

Knowledge of "cause-effect" provides the basis for decision making [30] and predictive analytics in particular.

Causal reasoning, as a "process of observing an event and reasoning about future events that might be caused by it" [31], can be extremely helpful in solving complex problems such as identification and prediction of a particular event, crime suspects, fraud cases, detection of trends, question answering, support of decision making by politicians, businessmen, and individual users. An important feature of causality is the continuity of the cause-effect connection.

Many efforts have been made to extract "cause-effect" relationships from texts utilizing constraints and machine learning techniques [25].

In spite of the existing algorithms, dealing with causal reasoning [26, 31, 32, 33], there isn't a reliable computer system that can process big data and show good results in giving answers to such questions as:

What may cause cancer?
• tobacco use
• alcohol use
• being overweight or obese
• unhealthy diet
• wireless devices
• stress
• light bulbs
• implants

What may help fight cancer?
• ultrasound
• peppers
• nanoparticles
• mushrooms
• broccoli sprouts

As a way for solving the problems we propose extraction and processing of “cause-effect” relations on the basis of semantic patterns described above.
Semantic patterns approach helps to extract information dealing with “cause-effect” in order to make predictions for decision making. The information could be valuable in many subject areas, including medicine, biology, science, technology, etc.
The constructed rules use causality connectors [34, 35, 36] such as “cause”, “due to”, “lead to”, “result from”, “result in”, “owing to”, “therefore”, “if-then constructions”, etc.
BuzzTalk submits information for predictive analytics after processing thousands of texts in the natural language.
For example:

Statement	Processing results
Wireless Devices May Cause Cancer	What may cause cancer? • wireless devices
Snowstorm Caused Flight Cancellation	What may cause flight cancellation? • snowstorm
Medical Bills Cause Most Bankruptcies	What may cause bankruptcies? • medical bills
Cancer is caused by accumulated damage to genes.	What causes cancer? • accumulated damage to genes
The company lost a great deal of money. Hence, the CEO was asked to resign	Why was the CEO asked to resign? • the company lost a great deal of money
The car didn't brake in time. The reason was that the road was slippery	Why didn’t the car brake in time? • the road was slippery
Scientists generally believe that the combustion of fossil fuels and other human activities are the primary reason for the increased concentration of carbon dioxide	What is the reason for the increased concentration of carbon dioxide? • combustion of fossil fuels and other human activities

“Cause-effect” is often the next step after the extraction of objects or events from texts.
“Cause-effect” relations help to reason about the detected events and is vitally important for problem solving.

4. Concusion

Processing of texts in a natural language necessitates the solution of the problem of extracting meaningful information from big data. Diverse information is of great importance for decision making on products, services, events, persons, industries, organizations. Semantic relations play

a major role in solving different problems ensuring interaction with the information in a natural way. Semantic relations ensure tracing of interrelated knowledge. Semantic knowledge modeling can answer diverse questions about persons, their motives and patterns of behavior.
Semantic patterns approach is proposed as a solution to the problem of processing big data. The approach can effectively capture, store, analyze, and present big data. The developed semantic-based linguistic platform has been successfully realized in BuzzTalk portal for opinion mining, mood state detection, event extraction, economic activities detection, subject domain recognition, named entity recognition and predictive analytics, thus helping to solve the problem of automated reasoning for decision making. The approach ensures high accuracy, flexibility for customization and future diverse applications for information extraction.
Implementation results show that the proposed knowledge-based approach (with statistical methods involved to prevent unwanted results) is correct and justified and the technique is highly effective.

5. References

1. Simon Ph. Too Big to Ignore: the Business Case for Big Data. Wiley, 2015. – 256 p.
2. Davenport Th. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press, 2014. – 228 p.
3. Mayer-Schönberger V., Cukier K. Big Data: a Revolution that will Transform How We Live, Work, and Think. Boston: Hough-ton Mifflin Harcourt, 2013. -- 242 p.
4. Marr B. Big Data - Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. Wiley, 2015. -- 256 p.
5. Marr B. Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things. Kogan Page, 2017. -- 200 p.
6. Moens M. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer, 2006. - 246 p.
7. Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval: the Concepts and Technology behind Search. Addison-Wesley Professional, 2011. - 944 p.
8. Buettcher S., Clarke C., Cormack G. Information Retrieval: Implementing and Evaluating Search Engines. MIT Press, 2010. - 632 p.
9. Machová K., Bednár P., Mach M. Various Approaches to Web Information Processing. Computing and Informatics, Vol. 26, 2007, p. 301–327.
10. Khoo Ch., Myaeng S. H. Identifying Semantic Relations in Text for Information Retrieval and Information Extraction. Springer, 2002. – p. 161- 180.
11. Bobkov A., Gafurov S., Krasnoproshin V., Romanchik V., Vissia H. Information Extraction Based on Semantic Patterns. Proceedings of the 12-th International Conference - PRIP'2014, Minsk, 2014, - p. 30-35.
12. Barnbrook G., Mason O., Krishnamurthy R. Collocation: Applications and Implications. Palgrave Macmillan UK, 2013. - 254 p.
13. Cruse D.A. Lexical Semantics. Cambridge University Press, 1986. - 310 p.
14. Manning C. D., Schütze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press, 1999. - 620 p.
15. Bilan V., Bobkov A., Gafurov S., Krasnoproshin V., van de Laar J., Vissia H. An Ontology-Based Approach to Opinion Mining. Proceedings of 10-th International Conference PRIP'2009, Minsk, 2009, - p. 257–259.
16. Fensel D. Foundations for the Web of Information and Services: A Review of 20 Years of Semantic Web Research. Springer, 2011. - 416 p.
17. http://www.buzztalkmonitor.com
18. Pang B., Lee L. Opinion Mining and Sentiment Analysis. Now Publishers Inc, 2008. – 148 p.
19. Devitt A., Ahmad K. Sentiment Analysis in Financial News: A Cohesion-based Approach //

Proceedings of the Association for Computational Linguistics (ACL' 2007). – p. 984–991.
20. Eguchi K., Lavrenko V. Sentiment retrieval using generative models // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP' 2006). – p. 345–354.
21. Clark A.V. Mood State and Health. - Nova Publishers, 2005. - 213 p.
22. McNair D.M, Lorr M., Droppleman L.F. Profile of Mood States. - San Diego, Calif.: Educational and Industrial Testing Service, 1971.
23. Siegel E. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Wiley, 2013. – 320 p.
24. Mishra N. Predictive Analytics: A Survey, Trends, Applications, Opportunities & Challenges, International Journal of Computer Science and Information Technologies, (vol. 3) (2012) - p. 4434- 4438.
25. Asghar N. Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey. arXiv preprint arXiv:1605.07895, May 2016.
26. Sorgente A. Automatic Extraction of Cause-Effect Relations in Natural Language Text. Proceedings of the 13th Conference of the Italian Association for Artificial Intelligence, 2013, - p. 37–48.
27. Darian S. Cause and effect in a corpus of science textbooks. ESP. Malaysia, 4, 1996. - p. 65-83.
28. Khoo C.S.G.. Automatic identification of causal relations in text and their use for improving precision in information retrieval. (Doctoral dissertation, Syracuse University, 1995). Dissertation Abstracts International, 5704A, 1364.
29. Xuelan F., Kennedy G. Expressing causation in written English, RELC Journal, 23(1) (1992). - p. 62-80.
30. Chan K., Lam W. Extracting Causation Knowledge from Natural Language Texts, International Journal of Intelligent Systems, vol. 20 (3) (2005). - p. 327–358.
31. Radinsky K., Davidovich S. Learning to Predict from Textual Data, Journal of Artificial Intelligence Research, (45) (2012). - p. 641-684.
32. Radinsky K. Learning Causality for News Events Prediction. Proceedings of the 21st International Conference on World Wide Web. ACM, 2012. - p. 909–918.
33. Kaplan R., Berry-Rogghe G. Knowledge-Based Acquisition of Causal Relationships in Text, Knowledge Acquisition (3) (1991). - p. 317-337.
34. Wolff P., Song G., Driscoll D. Models of Causation and Causal Verbs. Meeting of the Chicago Linguistics Society, main session, vol. 1, 2002, p. 607–622.
35. Levin B., Hovav M.A. Preliminary Analysis of Causative Verbs in English, Lingua (92), 1994. p. 35–77.
36. Altenberg B. Causal Linking in Spoken and Written English, Studia Linguistica, 38(1), 1984. p. 20-69.