Saint Petersburg State University

Graduate School of Management

Master in Management Program

USING POLARITY CLASSIFICATION MODEL TO ASSESS CUSTOMERS' ATTITUDES: THE CASE OF RUSSIAN E-COMMERCE COMPANIES ON TWITTER

Master's Thesis by the 2nd year student Concentration — Information Technologies and Innovation Management Alexander P. Timakov

Research advisor: Associate Professor, Sergey A. Yablonsky

Saint Petersburg

2018

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Я, Тимаков Александр Павлович, студент второго курса магистратуры направления «Менеджмент», заявляю, что в моей магистерской диссертации на тему «Использование модели определения тональности твитов для оценки отношения клиентов на примере российских интернет-магазинов», представленной в службу обеспечения программ магистратуры для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Мне известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

____ (Подпись студента)

25.05.2018

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF THE MASTER THESIS

I, Alexander Timakov, second year master student, program «Management», state that my master thesis on the topic «Using Polarity Classification Model to Assess Customers' Attitudes: the case of Russian E-Commerce companies on Twitter», which is presented to the Master Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from master theses, PhD and doctorate theses, which were defended earlier, have appropriate references.

I am aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St.Petersburg University «A master thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Professional Education Saint-Petersburg State University «a student can be expelled from St.Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».

____(Student's signature)

25.05.2018

ABSTRACT

Master Student's Name	Alexander P. Timakov
Master Thesis Title	Using Polarity Classification Model to Assess Customers' Attitudes: the case of Russian E-Commerce companies on Twitter
Faculty	Graduate School of Management
Main field of study	Information Technologies and Innovations Management
Year	2018
Academic Advisor's Name	Associate professor, Sergey A. Yablonsky
Description of the goal, tasks and main results	Russian E-Commerce companies fight with international players for customers and are interested in implementing Big Data tools for more advanced analysis of existing data upon their customers to get more profound understanding of their attitudes. Knowledge of Customers' Attitudes allows these companies to enhance their digital activities and obtain additional value. Polarity Classification allows obtaining knowledge upon sentiment and meaning of Customers' Attitudes from unstructured textual data in Social Networking Sites. The research goal of the thesis was to create and test Polarity Classification model, which allows managers of Russian E-Commerce companies to extract additional knowledge about Customers' Attitudes towards their companies from User-Generated Content. To achieve this goal, several Russian E-Commerce companies were selected as baseline examples (Wildberries, Citilink, M.Video, Lamoda, Sportmaster). Existing state-of-the-art Polarity Classification models were compared in terms of their performance (Machine Learning based models were the most efficient ones). Initial Polarity Classification models based on theoretical review were created and compared with existing services for Polarity Classification of Russian language (initial models showed higher accuracy). Finally, created model was applied to real-world testing tweet dataset (with indirect mentions of researched companies) and possible managerial applications of obtained 'Topic-Sentiment' knowledge for researched companies were discussed.
Keywords	Big Data, Polarity Classification, Social Networking Sites, Customer Attitude Assessment, Russian E-Commerce

АННОТАЦИЯ

Автор	Тимаков Александр Павлович
Название магистерской диссертации	Использование модели определения тональности твитов для оценки отношения клиентов на примере компаний российского рынка электронной коммерции
Факультет	Высшая Школа Менеджмента
Направление подготовки	Менеджмент инноваций и информационных технологий
Год	2018
Научный руководитель	Кандидат технических наук, доцент Яблонский Сергей Александрович
Описание цели, задач и основных результатов	На российском рынке электронной коммерции идёт борьба за клиентов между локальными компаниями и международными игроками. Компании на этом рынке проявляют заинтересованность во внедрении инструментов больших данных для более глубокого анализа существующих данных своих клиентов, чтобы получить понимание их мнений относительно компании и её продуктов. Знание отношения клиентов к определённым аспектам работы компании позволяет усилить маркетинговую деятельность и извлечь дополнительную ценность. Такой инструмент, как определение тональности, позволяет получить знания о настроениях и значении отношений клиентов из неструктурированных текстовых данных из социальных сетей. Целью исследования является создание и апробация модели определения тональности твитов, позволяющей менеджерам российских компаний рынка электронной коммерции извлекать из сгенерированного пользователями контента дополнительные знания об отношении клиентов к исследуемым компаниям. В качестве исследуемых компаний были выбраны несколько российских игроков рынка электронной коммерции (Wildberries, Citilink, M.Video, Lamoda, Sportmaster). Проведено сравнение существующих современных моделей определения тональности твитов с точки зрения их эффективности и аккуратности (построенные на машинном обучении модели оказались наиболее точными). На основе теоретического обзора были созданы первоначальные модели определения тональности, которые были сопоставлены с существующими сервисами для классификации полярности русского языка. Наконец, улучшенная модель была применена к реальному тестированию набора данных (твиты с косвенным упоминанием исследуемых компаний), и были обсуждены возможные управленческие приложения полученных в формате "Аспект-Тональность" знаний для исследуемых компаний.
Ключевые слова	Большие данные, анализ тональности текстов, социальные сети, оценка отношения клиентов, российская электронная коммерция

Table of contents

List of tables and figures

Tables

Figures

List of abbreviations

To avoid misunderstandings and misinterpretations, the list of abbreviations used in current thesis is provided below:

API — Application Programming Interface

AI — Artificial Intelligence

BD — Big Data

DB — Database

DL — Deep Learning

DNN — Deep Neural Networks

DM — Data Mining

KDD — Knowledge Discovery in Databases

ML — Machine Learning

NB — Naïve Bayes

NLP — Natural Language Processing

NLTK — Natural Language Tool Kit

NN — Neural Networks

SA — Sentiment Analysis

SNS — Social Networking Sites

SVM — Support Vector Machines

Introduction

E-Commerce is booming all around the globe, growing with double-digit rates (Statista, 2017). In Russia, online sales exceeded 1 trillion rubles and E-Commerce market has shown growth rate of 13% in 2017, even as offline retail was severely affected by the economic crisis (AITC, 2018). At the same time, volumes of cross-border trading is growing up faster, than local E-Commerce and several huge international players, such as Aliexpress, Asos and JD.com entered the market in the last 2 years (Data Insight, 2017). Along with increase in supply, there is 6% decline in consumers' purchasing power in 2016 (FSSS, 2017). These factors led to more intense competition for customers' attention and approval. To get this approval, company should identify and fulfill customers' demands and needs. The question every marketing and sales specialist in E-Commerce should ask is 'how to change customers' attitude in favor of our company'. Efficient communication between firms and customers can promote selling of particular products, services and brands. Thus, understanding consumers' attitudes is the core of creating and making marketing decisions (Sheng et al., 2017).

In E-Commerce, where all the company-customer interactions happen online and often in real-time, the rational venues to collect data upon customers' attitude are internal sources (sales data, marketing researches) or external sources, such as online forums and social networking sites (SNS). SNS are vital part of humans' daily lives and are widely spread around the globe (Statista, 2018). Such diffusion leads to creation of huge amounts of User-Generated Content (UGC), which in many cases contains customers' sentiments towards companies, their brands, products, services. UGC is online word-of-mouth behavior and is represented in form of unstructured data. Existence of such sort of information allows companies to forget about surveys, focus groups and external consultants to find consumer opinion about its products and those of its competitors (Liu, 2010). To collect and interpret this type of unstructured data is more efficient with application of Big Data (BD) techniques and tools.

'Big Data' is a buzzword of the beginning of 21st century. Numerous companies in different industries — from small fashion boutiques to multinational pharmaceutical conglomerates — utilize this technology trend to reach and verify their competitive advantage, since it positively affects both companies' strategy and operations (Hagen, 2013). It allows creating design-driven innovations and changing the paradigm of company-customer relationships (Morabito, 2015). BD proved to be especially useful from marketing perspective, since it allows companies to gather and analyze unstructured data and study consumer behavior and hidden consumer sentiment with help of it (Michael & Miller, 2013).

There is a wide variety of possible applications of BD extracted from SNS, which are beneficial to a company — company can predict adoption probability (Fang et al., 2013), improve consumer-retailer loyalty (Rapp et al., 2013), boost advertising and revenue growth (Shriver et al., 2013). Through analysis of company-related UGC, company is capable of identifying consumers' attitude towards its products and services (Pozzi, 2017). This sounds interesting to companies, because knowledge of customers' attitude allows company to tune its marketing strategy (including niche market identification and brand positioning) and interaction with customers (Bahtar & Muda, 2016), which directly influence end users' decision upon purchase of company's products and services (Ding et al., 2015). This is the reason, why having a flexible and powerful (in many cases, free) toolkit to leverage brand-related openly accessible UGC in favor of extracting knowledge about SNS users' attitude from optional feature became a 'must-have' (MongoDB White Paper, 2016).

This concurrent learning of users' behavior is beneficial to real-time, intent-based optimal interventions, which increases purchase likelihood (Ding et al., 2015). However, many E-Commerce companies in Russia do not even try to benefit from using this type of information in spite of its appealing possible outcomes, or most companies are capturing only a fraction of the potential value of data for the sake of improving its sales efforts (Tadviser, 2017). One of the reasons for that may be lack of theoretical base clearly aligning application of innovative BD techniques toward digital marketing benefits (Amado, 2018).

Since most of the data from SNS (online reviews, UGC, online ratings) is raw textual data, such toolkits as Natural Language Processing (NLP) may be applied. This frontier domain of BD and Artificial Intelligence (AI) is aimed at text extraction, preparation and analysis, and deals with human-computer language interaction (Devika et al., 2016). It is applied in such spheres, as spam filtering, search recommendations and chat bots. One of NLP subdomains — Sentiment Analysis (SA) — is specifically designed to work with attitudes within textual data (Pozzi, 2017).

How SA may be beneficial for business? SA of UGC allows extracting knowledge about customers' attitude, thus, to make efficient data-driven decisions upon brands digital marketing activities (Sheng et al., 2017; Rambocas & Pacheco, 2018). The implementation of developed NLP practices of this type will be beneficial for any type of companies. For E-Commerce companies, implementation of such SA types, as opinion mining or polarity classification, in marketing process may be applied for online evaluation of customer satisfaction, better understanding of consumers and market (Nassirtoussi et al., 2014).

Research on SA of English language is comprehensive and includes numerous studies upon various SA tasks (Devika et al., 2016; Mantyla et al., 2018). Research on SA of Russian language is more limited in its variety and is concentrated on studies upon sentiment lexicon generation (Klekovkina & Kotelnikov, 2012; Rubtsova, 2013; Rubtsova, 2015), opinion search and retrieval (Kravchenko, 2012) and polarity classification (Kotelnikov, 2012; Loukachevitch et al., 2015). However, when it comes to more business-oriented research with actionable outcomes of SA tasks in Russian language, amount of studies is very limited (Ermakov, 2009; Polyakov et al., 2012; Kirilenko & Stepchenkova, 2017). Moreover, relying on overview of 300+ articles and conference presentations on topic of SA of Russian language in last 7 years (Dialog-21, 2012-2017; ROMIP, 2010-2015; RUSSIR, 2016-2017), it is legit to claim that there is the absence of business-oriented research related to application of such SA tasks, as subjectivity and polarity classification, in E-Commerce companies. Due to the fact, that SA algorithm have been tailored to a specific language given the complexity of having a number of lexical variations and errors introduced by the people generating content (Tellez et al., 2017), research of applications of SA of English language in E-Commerce cannot be seamlessly applied to SA of Russian language in E-Commerce.

All of the abovementioned facts indicate that there is a research gap, which this master thesis will fulfill. **Research gap** is in the lack of empirical studies upon applications of polarity classification of Russian language that are beneficial for managers of E-Commerce companies.

To fulfill stated research gap, the following **research objectives** were formulated**:**

- To review what knowledge of CA may help E-Commerce companies and how it may be extracted from UGC on SNS;
- To review applicability of BD and DM approaches in UGC collection and analysis;
- To get acknowledged with NLP as a toolkit for textual data analysis;
- To get acknowledged with SA fundamentals, types (with focus on subjectivity and polarity classification), models and what value it brings to E-Commerce companies;
- To review cutting-edge studies upon SA of English and Russian languages along with multilingual SA (with focus upon polarity classification);
- To review research upon applications of polarity classification of English and Russian languages (with focus on applications in E-Commerce);
- To review different models of polarity classification of English and Russian language and how their performance baselines are measured;
- To identify the criteria of choice of the most efficient models of polarity classification applicable to Russian language and E-Commerce business.

The following **research goal** was stated**:**

- To create and test polarity classification model, which allows managers of Russian E-Commerce companies to extract additional knowledge about customers' attitudes towards their companies from user-generated content.

To achieve this goal, the following **research questions** were stated:

- **(RQ1)** What type of polarity labeled data may be useful for manager in Russian E-Commerce' company?
- **(RQ2)** Where to find relevant UGC and how to extract data upon Customer Attitude (which is needed to build polarity classification model) from it?
- **(RQ3)** What are performance baselines for state-of-the-art polarity classification for Russian, English and multilingual models?
- **(RQ4)** How accuracy of these polarity classification models is measured?
- **(RQ5)** What are the most relevant polarity classification of Russian language' models for current research goal?
- **(RQ6)** How accurate are existing proprietary services of polarity classification of Russian language?
- **(RQ7)** How to create tailored for researched companies' polarity classification of Russian language' model aimed on assessment of Customer Attitude in UGC from SNS?

The master thesis consists of three stages of research. The first stage is devoted to theoretical part, where impact of CA on E-Commerce, efficiency of BD and Data Mining (DM) techniques for unstructured data collection and analysis, and NLP as tool for unstructured textual analysis are reviewed. In addition, SA process and approaches towards it along with value brought to business by SA are overviewed. The second chapter consists of research methodology, process of obtaining business and data understanding, data preparation and information upon modelling. The third chapter includes creation, evaluation, iterative improvement and deployment of polarity classification of Russian language' models with highest expected accuracy.

The research is based on qualitative approach and the research method to be applied is secondary data analysis. Secondary data includes UGC (such as posts, comments, likes, reposts and users' reactions from SNS) and analytical data from online sources (open-source datasets, already combined word corpora on NLP). The theoretical research conducted is based on numerous sources, which include scientific articles, books, industry reports and conference papers. The sources for supporting review of existing literature and methodologies were found in the following databases: EBSCO, Elsevier, Emerald, JSTOR and ProQuest. Polarity classification as a main empirical research model is applied. Several research steps are fulfilled. First, business (business goal and expected managerial applications are stated) and data understanding (suitable SNS in terms of data needed is identified) are obtained. Second, relevant data is collected (grabbing raw data from SNS, storing in database, sending to analysis environment) and prepared (initial data cleaning followed by translation of data from unstructured to semi-structured format with general and domain-specific text processing). Third, SA models are analyzed and compared. Forth, the most relevant models to subjectivity and polarity classification are chosen. Fifth, those models are evaluated and iteratively improved (e.g. via expanding of word corpus or changing feature selection approach). Finally, chosen models are deployed and all the needed analyses with data are performed.

The main research presupposition is that the knowledge of Customer Attitude is crucial for success of Russian E-Commerce companies, and that this can be collected and utilized with the help of Sentiment Analysis tools and unstructured textual open data from Social Networking Sites. By being able to identify and influence customer attitude, company can drastically improve the efficiency of its marketing efforts, company-customer interactions and consequent sales volume. The expected findings from current research would be the following:

- Types of polarity labeled data that are useful for Russian E-Commerce' managers;
- The most relevant sources of UGC and methods to extract useful insights on target audience from collected UGC;
- Performance baselines of different state-of-the-art polarity classification for Russian, English and multilingual models;
- The most relevant accuracy measurement metrics for polarity classification models;
- Criteria to identification of the most appropriate for specific task and industry SA model;
- Performance baselines of different existing proprietary services for polarity classification;
- Steps to create, evaluate and iteratively improve tailored for specific task and industry polarity classification of Russian language' model.

CHAPTER 1. E-COMMERCE, BIG DATA AND SENTIMENT ANALYSIS

Literature review is concentrated on investigation of how Big Data Natural Language Processing' techniques, such as Sentiment Analysis, may help E-Commerce companies' managers to obtain and understand Customer Attitude towards companies' products and services. Lack of applicable literature upon managerial implications of SA and lack of business-oriented studies upon application of polarity classification (task of SA) of Russian language were identified.

It consists of three parts (see Figure 1). The first part starts with description of E-Commerce, its elements and its role in modern economy. Then, the role of CA in purchase behavior process along with its elements and ways to measure it are described. Importance of knowledge of CA for E-Commerce companies is reviewed. In second part, BD and DM as highly efficient approaches to collect and utilize CA knowledge are described. Specifically, DM methodology used in empirical part is reviewed. Then, specifics of UGC from SNS are described. Finally, research of BD applications in E-Commerce is overviewed. The third part of literature review touches upon topic of NLP. It starts with overview of NLP and its characteristics. Next, different NLP techniques and tools are described. Then, Sentiment Analysis as a method to extract knowledge from textual data and how it may be applied to data gathered from SNS will be reviewed. Value of SA for E-Commerce companies is analyzed. SA modelling process is decomposed and thoroughly reviewed. Finally, current studies upon SA of English and Russian languages along with their applications in E-Commerce are reviewed.

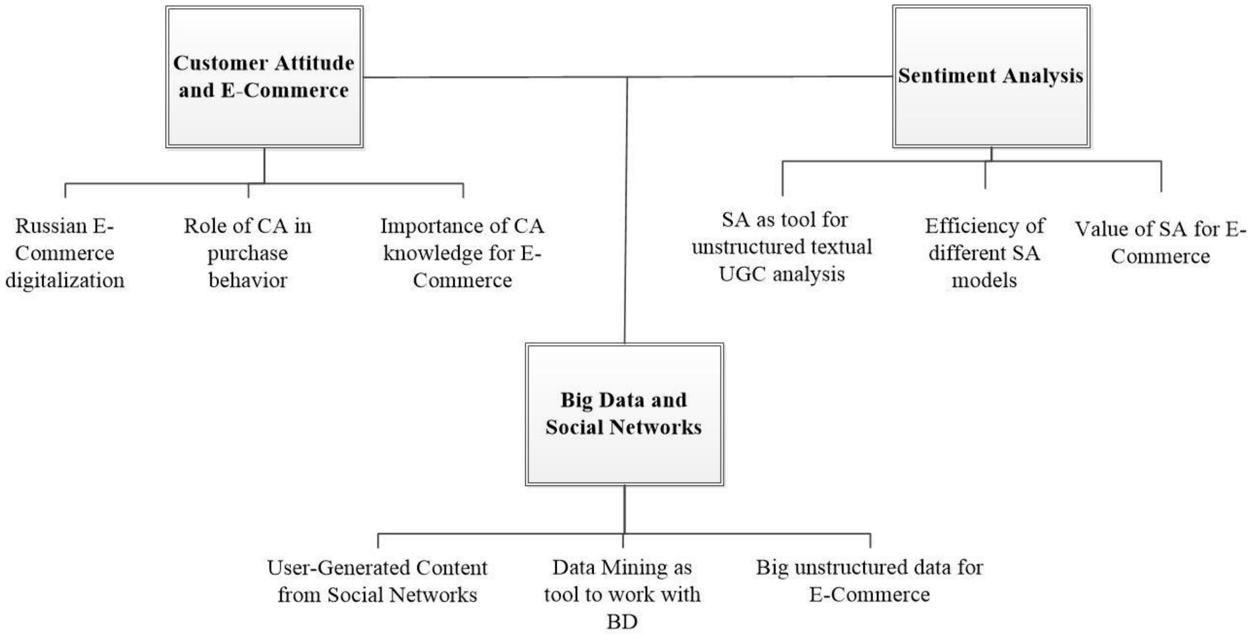

Customer Attitude and E-Commerce
Sentiment Analysis
Russian E-Commerce digitalization
Role of CA in purchase behavior
Importance of CA knowledge for E-Commerce
SA as tool for unstructured textual UGC analysis
Efficiency of different SA models
Value of SA for E-Commerce
Big Data and Social Networks
User-Generated Content from Social Networks
Data Mining as tool to work with BD
Big unstructured data for E-Commerce

Figure 1. Literature Review structure

1.1. How knowledge of Customer Attitude helps in E-Commerce

1.1.1. E-Commerce

The Internet has changed the nature of shopping in the past two decades, which has supported the proliferation of E-Commerce sites (Yadav & Rahman, 2017). E-Commerce relates to the use of electronic communications and digital information processing technology in business transactions to create, transform, and redefine relationships for value creation between or among organizations, and between organizations and individuals (Mohapatra, 2012). As one of the greatest technological developments in the last twenty years, E-Commerce has driven revolutionary change in global business, with the primary benefits including entrance into new markets, increased customer base, streamlined supply chains, improved customer service, increased profits and reduced costs (Karavdic & Gregory, 2005). Numerous examples of successful E-Commerce projects include Business-to-Customer (B2C) platforms (such as Amazon, Lamoda, Ozon), Business-to-Business (B2B) platforms (such as Alibaba, BizBilla, TragedHost) and Customer-to-Customer (C2C) platforms (such as eBay, Etsy, Craigslist). Although, along with feasible benefits, this fast growth of E-Commerce ecosystem led to increasing competition for customer all around the globe.

Growing number of Russian people prefer to purchase online (E-Commerce Foundation, 2016). Russian E-Commerce companies are in intense competition for customers (e.g., cross-border competitors such as China) (E-Commerce Foundation, 2016). To improve customers' satisfaction levels and brand comprehension, Russian E-Commerce companies need to enhance their digital marketing activities.

This significant structural shift towards online and digital retailing in recent years resulted in more personalized and communicative marketing approaches and created advantages for both the firm and consumer (Nisar, 2017). One of the reasons behind that rise of personalization of marketing approaches is in increasing ability of marketing practitioners to get more profound understanding of hidden customer needs and behavior patterns. Since customers' choice is strongly influenced by online reviews (Yoshida, 2018), getting better knowledge upon customer attitude from those online UGC is a starting point.

1.1.2. Customer Attitude

CA in current research is explained as customers' perception of some entity. It is one of the key element influencing final purchase decision within any model of customer engagement, satisfaction and purchasing behavior process (Lantos, 2010).

To see how attitude towards company affects customer decisions, general understanding of how purchase process goes was obtained. There are several common customer behavior models used in modern literature, such as Howarth-Sheth model and Engel-Blackwell-Kollat model. The former model (see Figure 2) constitutes customer behavior process from such stages, as information search (attention + stimulus + biases), purchase decision formation (motives, confidence and brand experience form attitude, which creates intention) and post-purchase behavior (loop of satisfaction-attitude-intention). It implies that customer post-purchase attitude is mainly affected by such aspects as customer attitude, brand comprehension and initial customer psychological constructs (confidence, motives, and purchase goals).

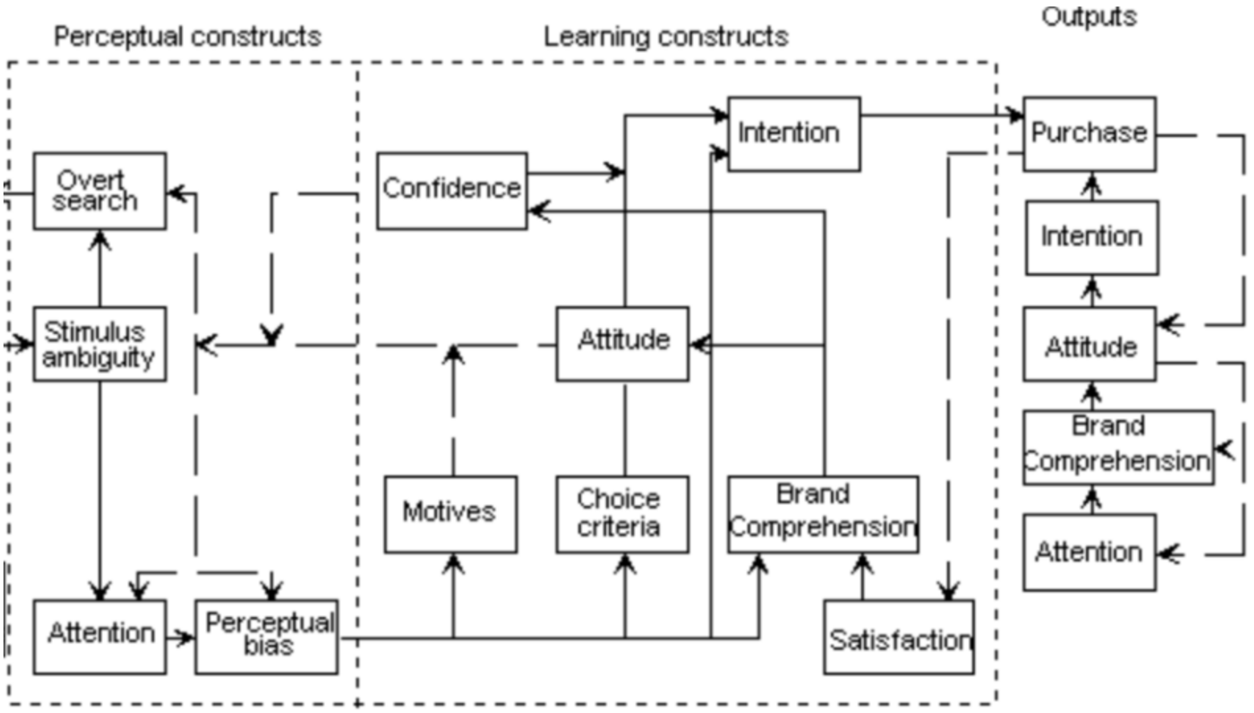
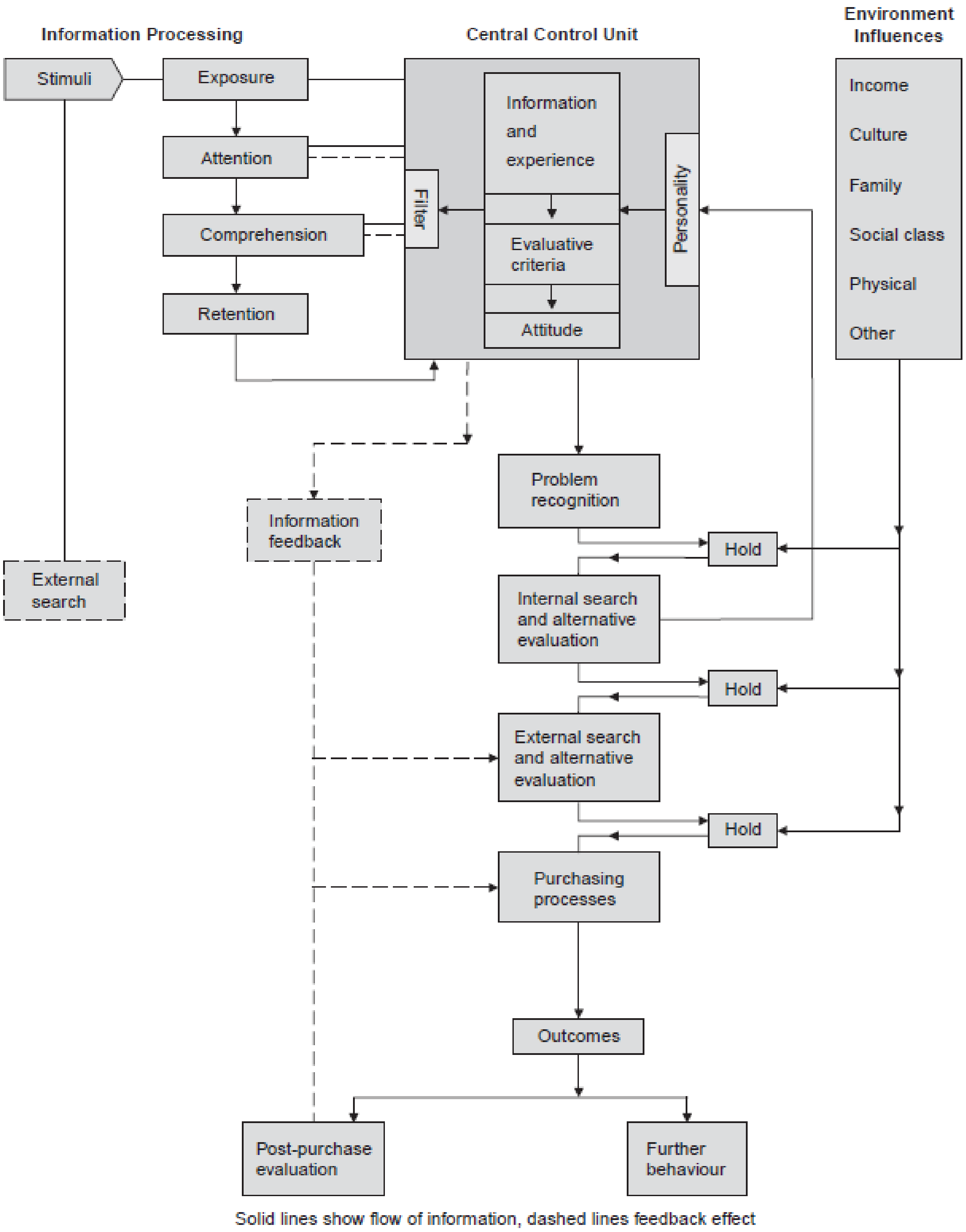


Figure 2. Howarth-Sheth customer behavior model
(Source: Howarth & Sheth, 1969)

The latter one (see Figure 3) implies more complicated process and includes various environment influences (such as income, social class, family and culture background). However, in this model customer's personality and brand comprehension are considered to be important criteria in the process of product evaluation and attitude building. It shows us how vital process of information feedback for customer's further behavior is.

Both models coherent in sense of similar logic of how customer come up with final purchase decision. After putting it in its simplified version, the whole process of product purchase can be divided into several essential parts: formation of customer attitude → evaluation of choices → initial purchase decision → customer satisfaction → customer post-purchase behavior with any sort of information feedback as a crucial indicator of customers' satisfaction and as an important affluence upon customers' attitude.



Solid lines show flow of information, dashed lines feedback effect

Figure 3. Engel-Blackwell-Kollat customer behavior model
(Source: Engel et al., 1979)

Thus, knowledge of customers' attitudes will allow companies to make their digital commerce activities more efficient (Hariharan, 2018; Wang, 2018). Traditional approaches for studying customer behavior require a large amount of time and resources. Existing research on customer behavior (including customer attitude, customer satisfaction and brand comprehension) typically relies on interviewing respondents with traditional paper-and-pencil surveys and online questionnaires. BD, NLP and open sources of relevant UGC allow making this process cheaper and faster (Song et al., 2016).

1.2. Big Data from Social Networking Sites

1.2.1. Big Data

What is Big Data? Half a century after computers entered mainstream society, the data has begun to accumulate to this threshold where something new and special is taking place. Now we live in a world where data is collected in ever-increasing amounts, summarizing more of what people and machines do, and capturing finer granularity of their behavior. Even if organization does not know exactly what it will do with this data, it can clearly perceive the value, which can be extracted from it. The quantitative change has led to a qualitative one and, considering an abundance of inexpensive computer power and enormous data amounts, this change is quite actionable. The sciences like astronomy and genomics, which first experienced the explosion of enormous amounts of raw unstructured data in the 2000s, coined the term ‘Big Data’ (Mayer-Schönberger, 2014). The concept is now migrating to all areas of human endeavor, including marketing, language processing and gathering vast volumes of information from different open sources, such as online forums and social networks.

The definition of BD varies greatly from publication to publication. The literature review has demonstrated that ‘big data’ term is applicable to a variety of different entities including social phenomenon, information assets, data sets, analytical techniques, storage technologies, processes and infrastructures. In the age of digitalization and datafication, information and the capability to extract useful knowledge from data is one of the company’s key strategic assets and solid foundation of new business models. Many definitions focus on characteristics of the data and on the differences between normal data and ‘Big Data’; the main question arising is ‘how big is big’. One way to think about it is that data is ‘big’ when we cannot fit it on one machine, another way — the data is ‘big’ when it goes beyond traditional limits in four major dimensions: volume, variety, velocity and value (Schutt, 2014). While it is problematic to generalize increase in value, changes in three other dimensions are measurable. According to one of the most famous frameworks, the 3-dimensional increase in data volume, velocity and variety invokes the need for new formal practices that will imply tradeoffs and architectural solutions that affect application portfolios and business strategy decisions. This ‘3 V’s’ framework is associated to the concept of ‘Big Data’ and used as its definition. Many other authors extended the ‘3 V’s’ model and, as a result, multiple features of big data such as value and veracity were later added to the list. Along with this definition, there are several other definitions of ‘Big Data’, which are dedicated to the crossing of some sort of threshold. For example, some researchers believe that data is big when it exceeds the processing capacity of traditional processing tools, such as conventional database systems and cannot be processed in an acceptable period or within a reasonable cost range (or requires the choice of an alternative way to process it).

Industry-wide approach to BD classification imply availability of three major dimensions, which help researchers to label concrete piece of data as 'Big Data' (Laney, 2001) are described in Figure 4:

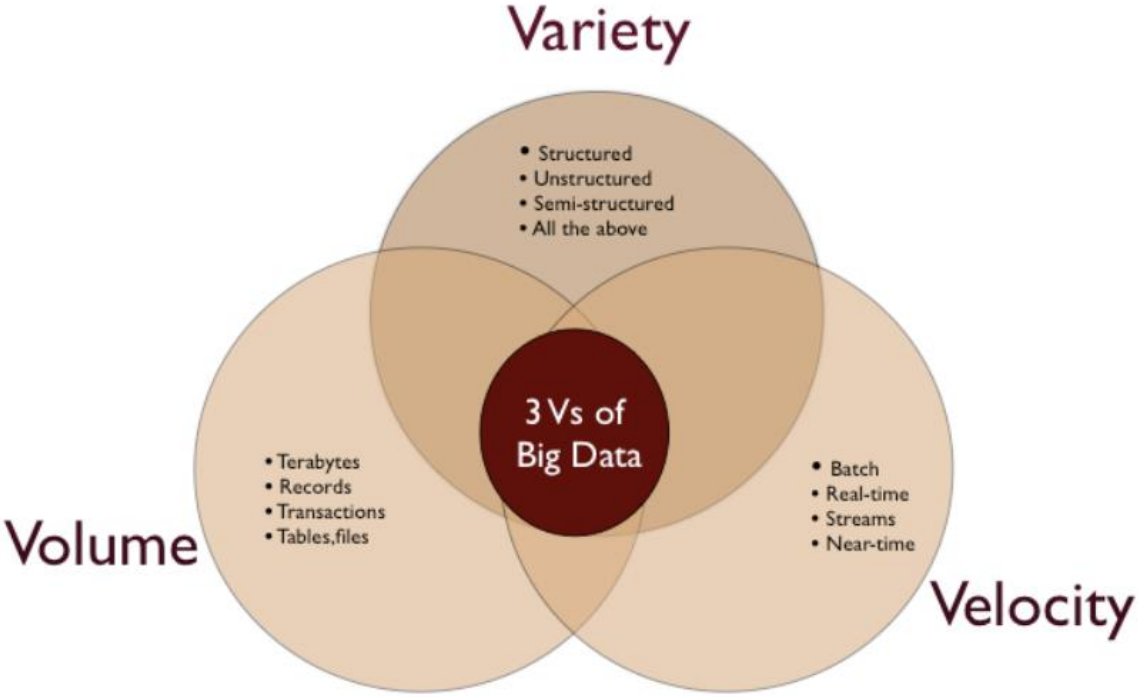


Figure 4. 3V's of Big Data model
(Source: Laney, 2001)

1. Volume, which is measured by the quantity of variables, transactions, number of data rows, attributes, interconnections and events. In the past researchers used to work mostly with samples — randomly (or based on an educated guess) chosen from the whole population small data sets, and created predictive models. However, big data does not assume any volume constraints, which allows researchers to analyze much larger data sets and identify a number of previously invisible trends and patterns. Today, volume is the defining factor when labeling some data as 'big data', because currently we are in Big Data 1.0 era, which is characterized by firms being busy with building capabilities and creating infrastructure to warehouse and process large amounts of data (Provost, 2013).

2. Variety, which represents the assortment of data and is closely connected with the definitions of structured, semi-structured and unstructured data. Structured data used to dominate

the amount of data being processed by enterprises and is much easier to analyze as it is classified on the basis of the data type (numeric, float, character, categorical, etc.) and usually is pre-processed (clean data with no missing values). However, over the past decades unstructured data has become the prevailing type of data for business analysts to work with. As companies started to look beyond organizational borders and expanded traditional operational data analysis, which most often comes in a form of structured data, they encountered a lot of unstructured, more complex data. Unstructured data by definition does not fit existing databases and is usually text heavy, yet may contain numbers and dates as well.

3. Velocity of data defines the speed and frequency of data creation, accumulation and processing. The pace of today's business world requires businesses to be able to perform real-time analysis on its KPIs and make appropriate decisions in real time. In this sphere big data opportunities are enormously various and impressive — from creation of parallel data warehousing to manage the pipeline of upcoming data to modelling comprehensive model with automatic visualization and reporting to decision-makers in company.

The amount of data in the world is doubling every two years and more than 95% of all this data is raw and unstructured, including images, videos, records, geo-based location data, network, and sensor data (McKinsey & Company, 2016). Therefore, one of the biggest obstacles for Big Data analytics in future is to gain an ability to master analysis of tons of unstructured data with ongoing application of meaningful results in practice. One of the approaches aimed on fulfilling this task is called 'data mining' (see Figure 5).

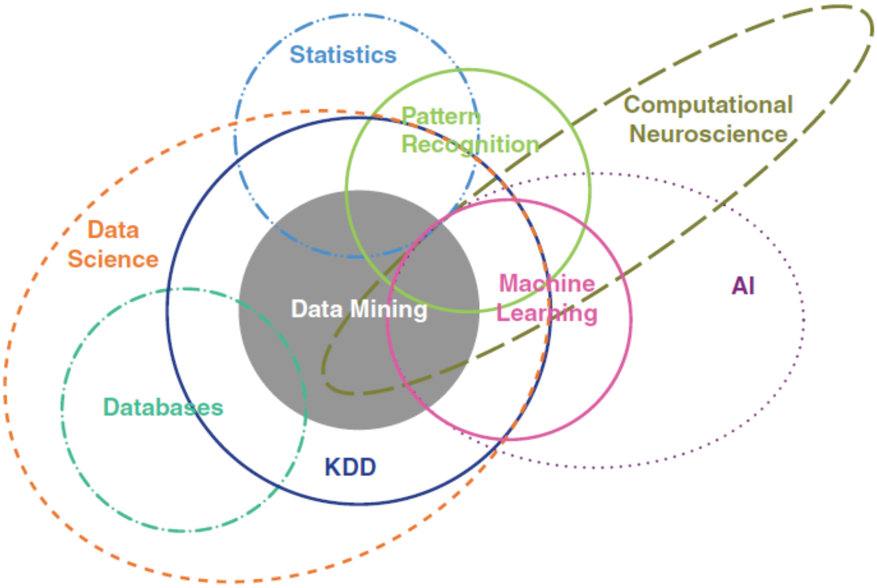


Figure 5. Multidisciplinary nature of Data Mining
(Source: Dean, 2014)

The most commonly accepted description of data mining is a complex process of extracting unknown, valuable knowledge such as pattern and mode from massive data. Besides, some refer to data mining process as knowledge discovery in databases, also known in data scientists' circles as KDD. There is an important distinction between data mining and KDD — data mining is usually focusing in the discovery part within KDD, which aims to find untouched information that holds potential value and build models for predictions. The results of data mining can be used in KDD in order to extract knowledge of the real case. Considering that the pattern discovered in data should serve the purpose of solving real questions, the use of data mining results should always influence and inform the data mining process itself (Provost, 2013). Data mining is being applied in different areas, and the detailed process differs in accordance with the task. The data mining process is interactive and iterative, involving numerous steps with many decisions made by the user.

There are several methodologies used to build the mining models. One of the most widely employed in industry methodologies is tool-neutral Cross Industry Standard Process for Data Mining methodology, which may be commonly separated into six key steps (Chapman et al., 1999):

1. Business Understanding. The most important notion on data analytics many specialists forget about is that all the tools and techniques provided by data science are aimed to help organization in making data-driven decisions, which will positively affect the value created by company and delivered to customers. Every analytics procedure should have some business-related purpose and not to be just an analysis in sake of analysis. It is vital to understand the business problem to be able to solve it in a data-driven manner, meaning that data analyst should think carefully about the use scenario of the results of her work and design a program of achieving stated business goals with regard of stakeholders' interests and historical knowledge. Final goal is to describe initial business goal (business parameter needed to be improved or altered in specific fashion) and additional actions (in a number of cases, output data may be used not only to achieve initial business goal).

2. Data Understanding. Prior to launching any collection or modelling process, researcher should clearly formulate output she want to get out of it, as well as get acknowledged with amount, quality and specifics of available relevant data. First, researcher should obtain understanding of input (data sources, criteria and formats suitable to business goals and available to researcher) and output (needed type of knowledge gained from input, such as patterns, groups, trends, etc.), as well as understanding upon data mining method (supervised/unsupervised). Before starting to process the collected data, analyst should get analyze unrefined data to get insights and find subsets valid

to represent the data. Final goal is to choose target data set (or focus on a subset of variables, or data samples) on which discovery is to be performed. Only after getting comprehensive business and data understanding analyst should switch to next data mining steps.

3. Data Preparation. Data cleaning and preprocessing is held on this stage. This may include removing of noise or outliers, collecting necessary information to model or account for noise and developing strategies for handling missing data fields or accounting for time sequence information and known changes. Final goal is to construct the tidy dataset applicable to be fed to modeling tools (Chiang, 2017).

4. Modeling. On this stage, researcher chooses upon the needed data mining algorithm. There are numerous specific algorithms, developed in recent years, but the number of fundamental ones may be limited to nine: classification and class probability estimation, regression analysis, similarity matching, clustering, co-occurrence grouping, profiling, link prediction, data reduction, and causal modeling (Provost, 2013). Out of all abovementioned data mining functions, we will be especially interested in regression analysis and classification. Next step is to build the data mining method with relevant chosen algorithm.

5. Evaluation. Final goal is to verify if created model is sufficient to achieve stated objectives (Chiang, 2017).

6. Deployment. On the final stage mining the data to find valuable knowledge and insights via identifying and interpreting observed patterns. Final goal is to launch the data mining process (immediate result) and to incorporate obtained knowledge directly or to carry out further analysis on it (further managerial implications).

In current case, to build sufficient data mining model, huge amount of textual data is needed. Since, nowadays is the era of Web 2.0, tons of personal unbiased information (including reviews, feedback, comments, etc.) may be found in increasingly accepted by public online social networks.

1.2.2. Social Networks

Online Social Networks are the most rapidly growing information sources which makes available an unprecedented scale of personal data, data about events and social relationships, public sentiments and behaviors that when are mined and interpreted are of an enormous value. Online Social Network (OSN) is a contemporary type of network whose history is relatively short but turbulent. The advent of mass adoption of online Social Networking Sites (SNS) has caused a shift on how people communicate and share knowledge, how businesses operate and compete and how politicians contest and influence. In the businesses field, social network analysis is applied to gain insight into markets and communities, with the "social enterprise" being the new necessity in order to manage knowledge, improvement, change, cooperation and risk. Many firms have embraced social media as a means to engage with their customers. Recent business reports have suggested that total spending on social media advertising has increased worldwide ($17.74 billion in 2014 vs. $11.36 billion in 2013, which amounts to an increase of 56.2%) and that social media engagement drives sales (Kumar, 2016). Online reviews are becoming increasingly important sources of information for shoppers influencing as much as 20-50% online purchase decisions (Mosteller, 2016), as well as an important source for companies analyzing users' demands and complaints. Better understanding of CA is seemingly one of the best tools that allows firm to identify and utilize points of growth in social media.

However, social network data are voluminous, even from a single social network site, mostly unstructured and their dynamic nature is evolving at an extremely fast pace that hinder the data analysis and extraction of knowledge. Having shifted away from the analysis of single small graphs and the properties of individual nodes to consideration of large-scale properties of graphs, the need for new data analysis tools and techniques is inevitable (Sapountzi, 2016).

All the content created and published by SNS users has its own name — User-Generated Content. UGC posted at SNS usually is in format of short messages with noisy (poor vocabulary, spelling and syntax) content. UGC fit for data mining and analysis tasks is based on several features: UGC is rich in embedded semantics and is dynamical (convenient in capturing attitude volatility over time). In addition, UGC has a positive effect on consumer's online product purchase intention, including increase in perceived credibility and usefulness of addressed entity, which lead to reciprocal positive customers' attitude towards UGC and influence their buying intention (Bahtar & Muda, 2016).

There are three main ways to access information in SNS:

- Application Programming Interfaces (APIs) (allows user to access website internal functionality from our source code);
- Crawlers / Scrapers (special applications aimed on grabbing data from non-protected websites);
- Direct request to SNS administration.

The format of extracted UGC data depends on way to access information and on tool used for this task. The most popular formats are .json, .csv, .txt and .xml. All of them are stored (e.g., in SQL/NoSQL databases) and processed in suitable for further analysis format (in case of current research — in .csv). There is great variety of tools to extract the UGC data and their analysis is out of current research scope. In current case, data collection was pipelined with the help of Node-Red and data formatting was performed with Python script.

1.2.3. Big Data in E-Commerce

Relevant for current research studies on applications of BD in business flourish and include such topics, as business analytics, business intelligence, web analytics, database management and numerous others. From organizational perspective, the most researched topics are organization strategy, organization and HR management with business intelligence tools and their integration as the main point of research. From operational perspective, operation management and supply chain management are in the center of attention. From marketing perspective, consumer behavior, consumer sentiment and marketing strategy are in focus of modern researchers. Such related to E-Commerce domains, as online communities (6 studies), sentiment analysis (23 studies), online ratings (7 studies), online reviews (28 studies) and electronic word-of-mouth (15 studies) are intensively researched (Sheng et al., 2017).

Briefly, any practical application of BD in E-Commerce (or Social Commerce) comes down to collection and analysis of some dataset — in a real-time or with some historical data. There are numerous ways to perform ongoing BD analysis on preprocessed dataset, which depends on sphere of further appliance and types of data collected (text, image, video, GPS tags, time logs, etc.). Two types of data analysis are generally used to extract needed for Social Commerce type of knowledge from datasets — Social Network Analysis (Link Prediction, Community Detection, and Influence Analysis) and Big Data Analysis (Trend Detection, Text Mining, and Collaborative Recommendation). Each of these types include various techniques, such as Machine Learning, classification estimation, clustering and regression analysis, each of which is aimed to solve specific set of tasks. In most of the studies, researchers used qualitative techniques and tried to use numerical data analysis, since direct usage of the raw textual data to extract customers' true opinions is a challenging task (Aguwa, 2017). However, the largest amount of collected data will be represented in text format, and one of the most useful for textual data analysis' tools is abovementioned text mining, and especially one of its part — Natural Language Processing (with Sentiment Analysis as one of its key parts).

1.3. Extract and Assess Customers' Attitudes with NLP and SA

1.3.1. NLP overview

Text mining or text analytics is a discipline that combines language science and computer science with statistical and machine learning techniques. It is used for analyzing texts and turning them into a more structured form to derive insights from it (Mount, 2016). It may be used for numerous tasks — from plain information retrieval to entities extraction and text classification with sarcasm detection.

Natural Language Processing is an attempt to extract a fuller meaning representation from free text. NLP typically makes use of linguistic concepts such as part-of-speech (noun, verb, adjective, etc.) and grammatical structure, either represented as phrases like noun phrase or prepositional phrase, or dependency relations like subject-of or object-of (Kao & Poteet, 2006). It is named 'natural' since it does not work with non-natural languages, such as Morse code or Klingon. It is one of the types of text mining — and it is highly relevant today, when it became a powerful machine learning backed instrument.

Machine learning (ML) is a discipline in Artificial Intelligence (AI) concerned with the design and development of algorithms that allow computers to reason and make decisions based on data (Swamynathan, 2017). In case of NLP, ML is a toolset of algorithms created to solve classification tasks, rather than radically different approach to solve data mining issues. Since NLP is built upon both statistics and ML, there are numerous interferences between those two groups' contributions to NLP and data mining in general.

ML in NLP is applied mainly as a set of classifier algorithms, such as Naïve Bayes, Support Vector Machines (SVM) and Maximum Entropy being most widely applied. Naïve Bayes is a sample probabilistic classifier, which shows more accurate results on small datasets. To define the final class of testing data, it counts the probability of sentiment to occur with regard to sentence it is located at and use Naïve Bayes formula to do that. SVM is a non-probabilistic classifier with large amount of training data required. It separates all testing points into k-dimensional hyper plane with non-linear function and after transform all objects into linearly separable map. Maximum Entropy is a conditional exponential classifier, which combines several features and uses their sum as exponent.

One of the subdomains of ML — Deep Learning — is especially relevant for modern NLP and brought a great benefit to it with the resurgence of Deep Neural Networks (Yin et al, 2017). Usage of such DL tools as Neural Networks brings sufficient improvement in accuracy for existing NLP techniques. Such algorithms, as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are widely applied in current research. RNN consist of one node, where all the input is fed up. While training, the model goes through number of epochs (one epoch = one forward + one backward propagation) and identify patterns (which are represented in form of matrices). There are two prevailing RNN types — long short-term memory (LSTM) and gated recurrent unit (GRU). CNN is a more complex construct, which consists of three layers — input layer (same as in RNN), convolutional layer (used for representational learning) and pooling layer. Choice of appropriate ML classifier, as well as choice of the most suitable DNN strongly depends on the task and type of input data, so there is no consensus on DNN selection for any particular NLP problem (Yin et al., 2017). However, both RNN and CNN in recent years have shown to perform very well in NLP tasks (Kalchbrenner et al., 2014).

1.3.2. NLP techniques

Development of NLP starts with document-level, and then go to sentence-level and aspect-level, currently stopping at morpheme-level. The final goal of NLP depends on further application of extracted data and on specifics of input/output data — it may be speech processing, text generation, information retrieval or others. Different NLP techniques are applied for different situations. However, all of them are based on usage of some parsers, needed to translate natural language to readable by machine binary language. Some parsers may be called off-the-shelf parsers, which uses open source morphological, semantical and syntactical libraries (MINIPAR, WordNet, VerbNet), and parsers that create their own libraries with the help of such techniques, as topic modelling (Kao & Poteet, 2006). The ongoing operations usually include data processing, analysis itself and final wanted by user output.

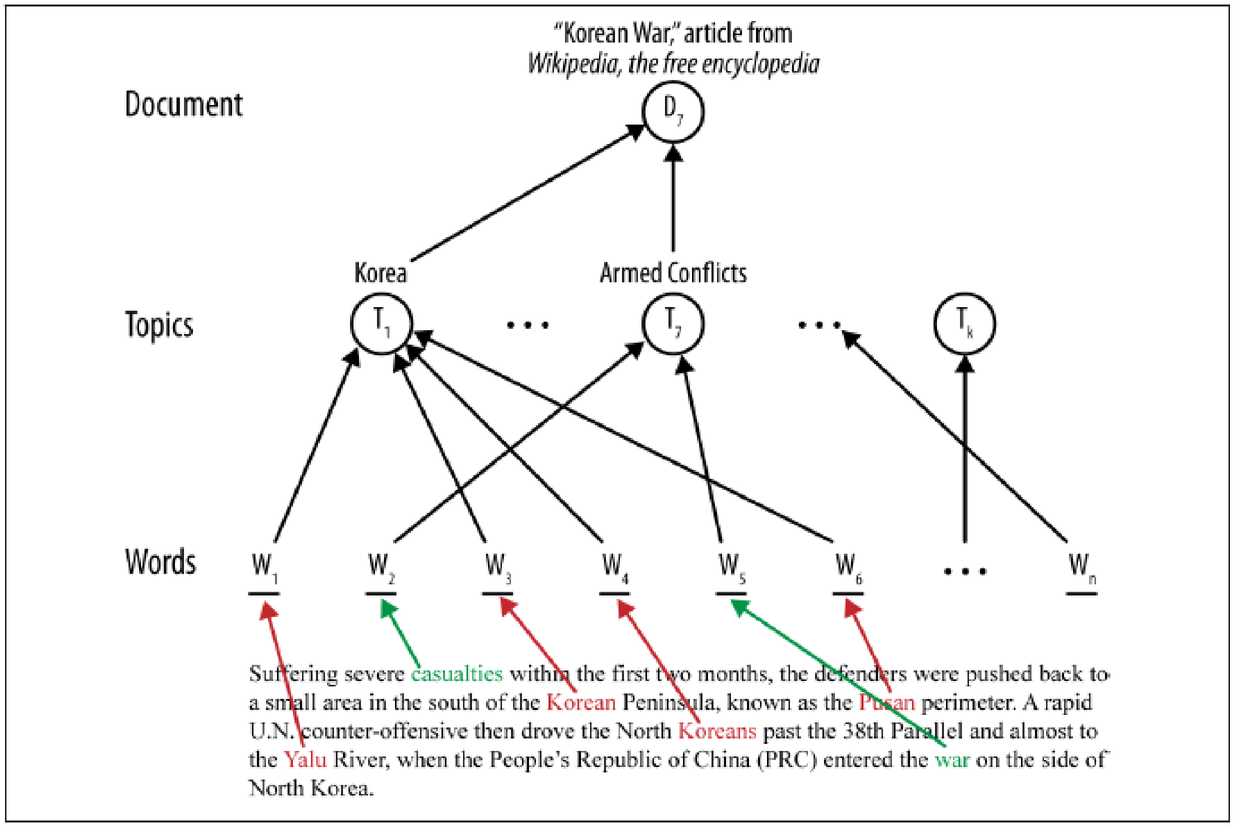

"Korean War," article from
Wikipedia, the free encyclopedia
Document
D_7
Korea
Armed Conflicts
Topics
T_1
T_7
T_k
Words
W_1
W_2
W_3
W_4
W_5
W_6
W_n
Suffering severe casualties within the first two months, the defenders were pushed back to a small area in the south of the Korean Peninsula, known as the Pusan perimeter. A rapid U.N. counter-offensive then drove the North Koreans past the 38th Parallel and almost to the Yalu River, when the People's Republic of China (PRC) entered the war on the side of North Korea.

Figure 6. Topic modelling example
(Source: Provost, 2013)

Technological stack used for Natural Language Processing is enhancing in recent years. It includes such open source tools as Python Natural Language Tool Kit, Stanford NLP, WordNet and FrameNet libraries, gensim and textblob modules and existing proprietary services, as Google Cloud based NLP service and IBM Cloud (backed by IBM Watson and supported by numerous applications, such as Cloud Floundry Apps and Tone Analyzer). Using any of these tools, it is possible to create a powerful application, which will analyze customers' attitudes from SNS towards companies' products and services. It may provide plain data or interactive visualizations as outcome, work by request or update in real-time — all these options depends on business needs of decision maker. Based on those analysis results, it is possible to come up with managerial solutions aimed on increase in number of satisfied customers, increase in overall number of users, increase in average revenue per user and decrease in churn rate for companies' products.

In combination with ML and DL, NLP is applied for various purposes — it allows performing machine translation, supporting spoken dialog systems, extracting product features from customer reviews and doing sentiment analysis. In modern business, NLP is used in numerous industries:

- Autocomplete and spelling correctors (based on named entity recognition) are integrated into search engines (Google, Yandex) and smartphone operational systems (Android, iOS);
- Automatic reasoning engines driven by natural language queries, such as Wolfram Alpha, helps to solve human's tasks;
- Google's 'Talk to Books' service, which exploits AI to perform advanced NLP and find proper literature recommendations based on one input phrase;
- Chatbots, separate or integrated into messengers — all of them use Speech-to-Text and Text-to-Speech NLP abilities to be able simply to input and output information;
- Sentiment Analysis tasks, such as polarity detection and opinion mining of UGC, are widely applied in finance industry (e.g., stock price prediction) and marketing (e.g., brand health monitoring).

1.3.3. SA overview

Sentiment Analysis has been one of the most active research areas in natural language processing since early 2000 (Liu, 2012). The aim of sentiment is to define automatic tools able to extract subjective information from texts in natural language and to create structured and actionable knowledge to be used by decision maker (Pozzi, 2017). As well as NLP itself, SA allows to work on different levels of granularity — from document-level to sentence and phrase level tasks (Agarwal, 2011).

Initially, it was regarded as standard document classification into topics (Pang, 2002) and since its early years required extended external sentiment lexicon (Nakov, 2017). Later, it became to be widely applied to more fine-grained words/phrases sentiment evaluation (Pang, 2005), but until rise of social media SA was either genre-agnostic (Baccianella, 2010), or focused on newswire texts (Wilson et al., 2005) and customer reviews from web forums (Pang, 2002; Pontiki, 2014). It was recognized that in many cases it is also crucial to know the topic toward which the sentiment is expressed (Stoyanov & Cardie, 2008), the role of context in determining the sentiment orientation (Wilson et al., 2005) and understanding of the linguistic aspects of expressing opinions, evaluations and speculations (Wiebe et al., 2004).

Currently, SA consists of various tasks, with some of them aimed on textual data' collection and processing (lexicon generation, labeled texts collection, etc.) and some aimed on textual data analysis (opinion mining, polarity classification, etc.). Taxonomy of the most popular SA tasks is presented in Figure 7.

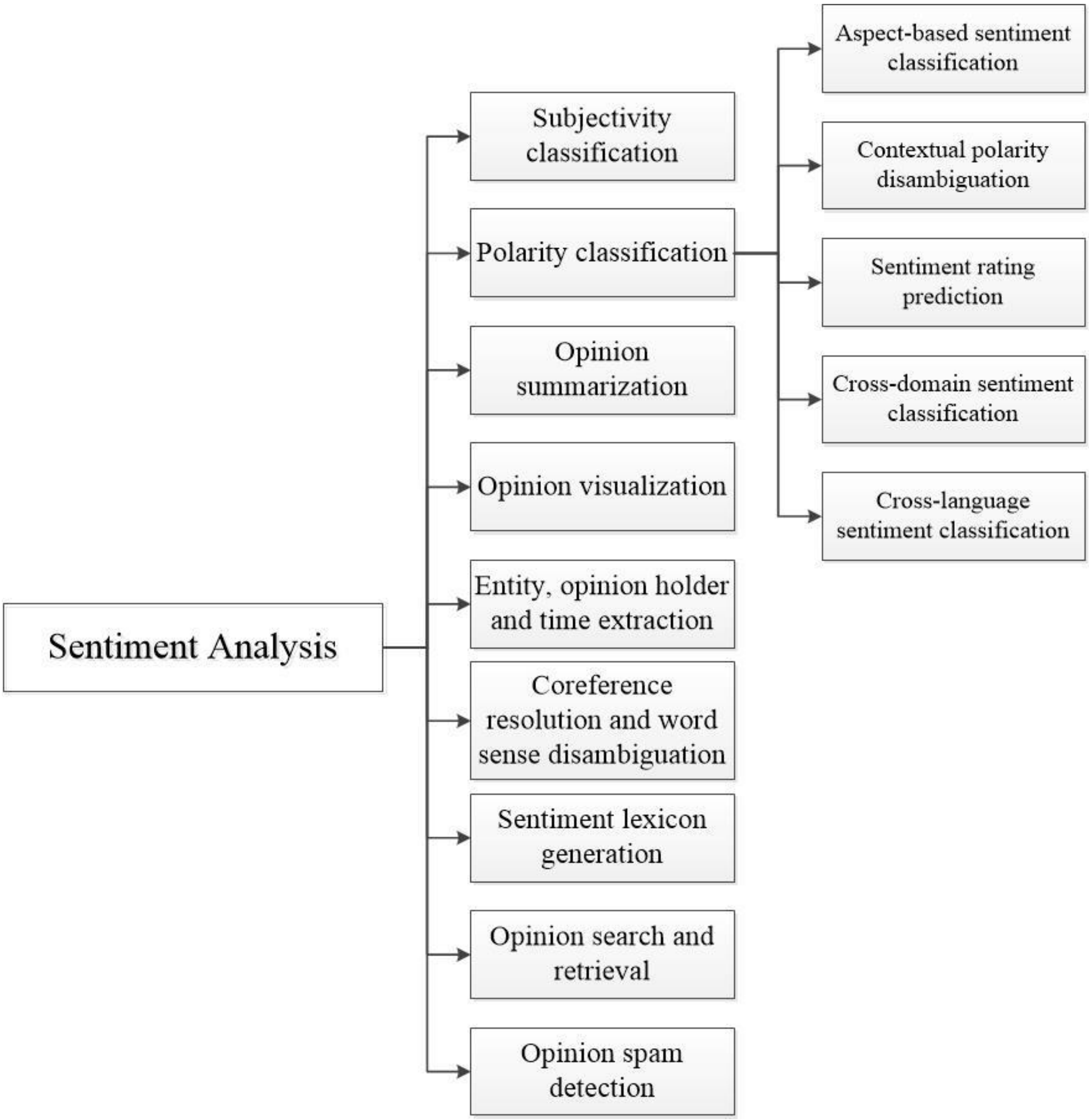

Sentiment Analysis
Subjectivity classification
Polarity classification
Opinion summarization
Opinion visualization
Entity, opinion holder and time extraction
Coreference resolution and word sense disambiguation
Sentiment lexicon generation
Opinion search and retrieval
Opinion spam detection
Aspect-based sentiment classification
Contextual polarity disambiguation
Sentiment rating prediction
Cross-domain sentiment classification
Cross-language sentiment classification

Figure 7. Taxonomy of the most popular sentiment analysis tasks (Source: Pozzi, 2017)

In the age of SNS, focus of SA studies shifted to research of UGC, which consists of natural language and has three key elements:

- Text — canonical language (semantics and punctuation), optionally with misspellings and punctuation missing;
- Part-of-Speech (POS) — shows if the text part is a noun, verb, or other language part. Some POS are indicators that helps to identify subjectivity (Pak & Paroubek, 2010) and polarity (Fersini, 2016);

- Paralinguistic Content (PC) — includes emoticons (😊), emphatic abbreviations (‘STFU’), onomatopoeic expressions (‘skrt’), lengthened words (‘daaaaaaamn’), capital letters (‘AWESOME’) and hashtags (‘#bund’).

All these types of data are exploited on different stages of modelling process and are typical to short-text online messages.

1.3.4. SA modelling process

Understanding the process and terminology of SA is fundamental for comprehension of further research overview and upcoming empirical modelling. To formalize the task, following definitions to be used:

- SA tasks — pool of various SA tasks, which includes such entities as subjectivity classification, polarity classification, sentiment lexicon generation;
- Subjectivity classification — performed to classify a sentence or a clause of the sentence as subjective (opinionated data) or objective (factual data) (Liu, 2010);
- Polarity classification — determine whether a sentence or a clause of the sentence expresses a positive or negative opinion;
- Holder — the entity (person / company) who holds opinion;
- Target — the entity (company / product / brand) addressed by the holder;
- Aspect — a part of target evaluated by opinion holder (e.g., in case of product as a target, aspects may be price, color, size, material, availability, etc.);
- Polarity — the sentiment towards target in general or some aspect specifically.

The most crucial issue for successful application of subjectivity and polarity classification is to obtain understanding of SA modelling process steps. Since modeling and data preparation steps are strongly interconnected, the only presupposition is that all the relevant textual data is collected and properly prepared before modelling steps.

There are various approaches to polarity classification, but starting point at creating appropriate model is in choosing level of analysis’ granularity. If researcher is interested in opinion about the whole entity within testing data, such as document or sentence, then SA modelling process consists from subjectivity classification and classifier creation. If researcher is interested in more precise evaluation and needs to identify both topic within sentence and sentiment towards it, then topic modeling techniques should be used and minimally viable version of SA modelling process consists of four steps: aspect extraction, feature selection, value assignment and classifier creation.

All the approaches to polarity classification specifically may be separated into two large groups — lexicon-based and ML-based approaches. Lexicon-based approaches may be based on usage of dictionary (dictionary-based models) or sets of rules (rule-based models). ML-based approach may be based on supervised learning (with ML algorithms as classifiers) or on unsupervised learning (with DL algorithms as classifiers). In addition, hybrid approaches (also called ensembles) with elements of both lexicon-based and ML-based approaches are used.

Along with those five approaches, some mixed approaches, which can not be categorized as lexicon-based, ML/DL-based or ensemble approaches due to their specific mathematical base, exist. For example, Formal Concept Analysis, Fuzzy Formal Concept Analysis and other concept-level sentiment analysis systems, such as pSenti and SenticNet (Medhat et al., 2014).

There are several steps, which are applied in different polarity classification models. One of them is feature selection. The main goal of feature selection is to select important for research aspects within sentences. Different approaches are usually applied for this, such as collection of bag-of-words, collection of n-grams, POS-tagging, binary occurrences and syntactic relations. In addition, another outcome researcher gets are quantifiable features (e.g., ratio of positive terms to the number of positive + negative terms, sum of all positive/negative terms, etc.).

Second of them is value assignment. Value assignment for lexicon-based models is sentiment assignment. Value assignment for models with no lexicon is weight assignment, which is a polarity vector evaluation for every element (which are selected earlier features). The main goal of weight assignment is to assign to each feature a proper label (numerous, since classifiers work only with numbers, not characters). Such techniques as TF-IDF, binary function and Pointwise Mutual Information (PMI) are used.

Third of them is classifier creation. In case of ML models, the most used classifiers are SVM, Naïve Bayes, Maximum Entropy and linear classifier. In case of DL models, the most used classifiers are Recurrent Neural Networks (especially with Long Short-Term Memory architecture) and Convolutional Neural Networks (CNN). The goal of classifier is to separate all the input data to several classes (predefined in case of supervised learning and identified in case of unsupervised learning). For ML models' classifiers would perform classification and regression tasks, while for DL models it would perform clustering task.

1.3.5. Research of SA

Every of SA process' steps are researched actively. Recently, hottest topics of cutting-edge research of SA include:

- Aspect-based SA (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016),
- SA of figurative language (Ghosh, 2015),
- Stance detection (Mohammad, 2016),
- Event polarity explication (Russo, 2015),
- Automatic lexicon building (Keshavarz & Abadeh, 2017).

As methods for SA mature, more attention is being paid to linguistic structure and to multi-linguality and cross-linguality (Nakov, 2017). Research of multilingual SA is rising and it includes:

- Development of specific lexicon (Steinberger et al., 2011);
- Supervised sentiment analysis in multilingual environments (Vilares et al., 2017);
- Sentiment analysis system adaptation for multilingual processing (Balahur & Perea-Ortega, 2015).

Research of Sentiment Analysis of Russian language is more focused on lexicon-related issues and includes:

- Part-of-Speech tagging (Anastasyev et al., 2017; Kazennikov, 2017);
- Morphological analysis (Selegey et al., 2016; Sorokin et al., 2017);
- Development of specific lexicons (Rubtsova, 2013; Rubtsova, 2015; Benko & Zakharov, 2016; Dubatovka et al., 2016; Kotelnikov et al., 2016; Mazurova, 2016);
- Ways of sentiment expression in Russian (Zagibalov et al., 2010);
- Opinion mining (Kravchenko, 2012).

Focus of current thesis will be allocated in the topics of subjectivity classification (for initial splitting of opinionated and factual data) and polarity classification. Objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs (Liu, 2010). Existing research of subjectivity classification and polarity classification of English language include:

- Understanding of word sense to disambiguate subjectivity (Wiebe & Mihalcea, 2006);
- Prior sentiment polarity of multi-word phrases (Russo, 2015);
- Aspect-based polarity detection (Pontiki et al., 2014).

Existing applications of subjectivity classification and polarity classification of Russian language are limited, since the most of research is based around lexicon generation and opinion mining, and include:

- Subjectivity vs. Objectivity dichotomy and sentiment relevance (Semina, 2018);
- Opinion retrieval from news articles (Chetviorkin & Loukachevitch, 2013);
- Entity-based sentiment polarity classification (Karpov et al., 2016);
- Overcoming problems of time gaps and data sparsity (Loukachevitch & Rubtsova, 2016).

Since sentiment analysis is of great importance to business and society, it has spread from pure computer science sphere to management science and the social sciences. Nowadays, it has gained even more value with the advent of social networks (Pozzi, 2017).

Existing research of Sentiment Analysis of English language for E-Commerce activities include:

- Discovery of consumer attitude insights via collecting and summarizing polarity upon product features (Chamlertwat, 2012);
- Measurement of marketing productivity on Twitter through sentiment and NPS assessment (Yoon, 2013);
- Discovery insights on more efficient resource allocation in online communities (Homburg, 2015).

Focus of our research will be allocated in the topics of subjectivity and polarity classification. Existing research of subjectivity and polarity classification of English language for E-Commerce activities include:

- Evaluation of customer satisfaction level using polarity towards a posteriori service aspects for parks and recreation sphere (Farhadloo et al., 2016);
- Identification of aspects within online reviews, which drive company’s product sales (Li et al., 2018).

Amount of research aimed on Sentiment Analysis of Russian language for E-Commerce activities is limited and it includes:

- Opinion extraction about cars (Ermakov, 2009);
- Decision support for quality management for hotel industry (Yussupova, 2012);
- Sentiment detection for banks and telecom companies (Arkhipenko et al., 2016; Karpov et al., 2016).

The main drawbacks of current business-oriented research towards polarity classification of Russian language are its pure theoretical orientation with no practical implications (e.g., the final goal of research is to compare different classifiers' accuracy on tweets' dataset with no further implications), or the lack of state-of-the-art approach upon solving business-oriented problems (e.g., using inflexible and labor-intensive lexicon-based models to solve specific business problems). There were found no research regarding state-of-the-art polarity classification of Russian language for E-Commerce with useful for decision-making recommendations, which makes this topic a research gap of current thesis. Neither purely theoretical, nor empirical studies were found, thus, this defines the research gap of current thesis.

1.3.6. Value of SA for E-Commerce

To come up with relevant applications of polarity classification of Russian language for E-Commerce, it is important to understand how Sentiment Analysis in general may be applied to business needs. Generally, all the currently researched applications may be divided into marketing-related and finance-related categories. Marketing-related applications allow managers to get better understanding upon such issues as:

- Influence of Producer-Generated Content on customers' conversion rates online (Ludwig et al., 2013);
- Influence of positive and negative UGC upon daily product sales (Sonnier et al., 2011);
- Factors within UGC and online reviews which affect customers' purchase decisions (Baek et al., 2012);
- Influence of UGC on new product adoption rates (Hennig-Thurau et al., 2015);
- Users' interest prediction based on opinions of their friends (Bao et al., 2013);

Finance-related applications allow managers to get better understanding upon such issues as:

- Impact of negative product reviews and ratings on volatility in returns and traded volumes (Tirunillai & Tellis, 2012);
- Stock price predictions based on Twitter mood (Bollen et al., 2012);
- Automated intraday stock recommendation system based on market data and textual news (Geva & Zahavi , 2013);
- Box office prediction based on microblog (Du et al., 2013).

Understanding of polarity towards companies' brands and products allows both middle and top-level managers within E-Commerce to enhance different activities, such as:

- To assess of customer satisfaction (Kang & Park, 2013),
- To make data-driven market decisions (Moreo et al., 2012),
- To predict and influence churn rates (Coussement & von den Poel, 2009),
- To improve recommendation systems (Garcia-Cumbreras et al., 2013),
- To adjust targeted advertising (Li & Shiu, 2012),
- To predict sales (Rui et al., 2013),
- To identify market trends and assess customer and partners' credibility (Li & Li, 2013);
- To predict market changes (Qiu et al., 2013).

Along with those activities, obtained with polarity classification and topic modeling information upon CA may be applied to improve not related to marketing and finance issues. Customers often mentions various aspects of companies' operational activities, such as logistics (e.g., time and quality of delivery, cost of delivery); IT infrastructure (e.g., stability of mobile app, relevance of content and response time of website); return policy (e.g., conditions for reclamations and refunds); consultants' service (e.g., how difficult it is to find a consultant, how competent she is); customer support (e.g., politeness and usefulness of provided help). It appears to be hard to bound the area of possible applications – UGC from Internet may include all sorts of customers' insights towards companies' brands, products, prices, service and even more subtle aspects, such as perception of companies' positioning and non-obvious competitive advantages.

1.3.7. Summary

NLP is widely applied technique for analysis of unstructured textual data. Along with ML and DL algorithms, it allows specialists manipulating with BD and create efficient models for such tasks, as speech recognition, information retrieval, machine translation and text processing.

SA is one of the fast-developing technologies of BD NLP, which allows studying customer attitude towards business entities (Pozzi, 2017). For example, it allows user to create a classifier (which is a machine-learning model concisely), train it with custom ready-to-update datasets and understand the intent behind text and returns a corresponding classification, complete with a confidence score. There are numerous research upon Sentiment Analysis in English (Chen & Zhang, 2014). At the same time, research on Sentiment Analysis in Russian is more focused on lexicon-related topics and is under researched in such tasks, as subjectivity and polarity classification (Chetviorkin & Loukachevitch, 2013). By identifying meaning and tonality of user-generated content (my goal and expected managerial applications), it is possible to understand what managerial decisions related to some product's or service's aspect should be implemented.

1.4. Research Gap

Having conducted comprehensive literature review of Big Data, Social Networks and E-Commerce, it was identified that proactive customer engagement and advanced digital marketing activities became crucial for success of E-Commerce company in age of Web 2.0 and Social Commerce. With the help of such technologies, as Natural Language Processing, Machine Learning and Sentiment Analysis, it is possible to get more comprehensive knowledge of customer needs and complaints.

From the literature review, it was determined that there are plenty of studies referring to Sentiment Analysis in English language. Due to the fact of high sensitivity of language algorithms, it is not possible simply to apply model, trained for English language, to Russian language — as well as apply model, trained on legal texts, to Twitter' short-text messages on the topic of electronic appliances. The lack of research on applicable to business needs Sentiment Analysis models in Russian language was identified. In addition, it was found that there are no studies on applications of polarity classification (a type of SA) for Russian E-Commerce companies.

With regard of research gap, several research objectives were stated. Topic of Customer Attitude and its impact on E-Commerce companies' marketing was reviewed and high importance of comprehensive understanding of CA was concluded. Topics of Big Data and Data Mining of User-Generated Content from Social Networking Sites were reviewed. Indispensability of Natural Language Processing for unstructured text analysis was identified. Fundamentals of Sentiment

Analysis were overviewed and such SA types, as subjectivity and polarity classification were found to be suitable tools to fulfill research goal. Final research goal is to build polarity classification of Russian language' model (as accurate as possible in comparison to state-of-the art performance baselines), which allows assessing users' attitude towards researched Russian E-Commerce companies by analyzing UGC towards company on chosen SNS. To fulfill this goal, seven research questions were stated:

- **(RQ1)** What type of polarity labeled data may be useful for manager in Russian E-Commerce' company?
- **(RQ2)** Where to find relevant UGC and how to extract data upon Customer Attitude (which is needed to build polarity classification model) from it?
- **(RQ3)** What are performance baselines for state-of-the-art polarity classification for Russian, English and multilingual models?
- **(RQ4)** How accuracy of these polarity classification models is measured?
- **(RQ5)** What are the most relevant polarity classification of Russian language' models for current research goal?
- **(RQ6)** How accurate are existing proprietary services of polarity classification of Russian language?
- **(RQ7)** How to create tailored for researched companies' polarity classification of Russian language' model aimed on assessment of Customer Attitude in UGC from SNS?

This research contributes to both theoretical and practical perspective. From practical perspective, if company (and its marketing specialists) has knowledge upon customers' attitude towards company's products, it can influence customer satisfaction level of its online community members (users of their products and services) and improve overall brand comprehension through fine-grained targeting, customized positioning and enhancing digital activities. In addition, the thesis is country specific since from theoretical part, because SA of Russian language is researched, and from practical perspective, since the final SA model is tailored to specifics of Russian E-Commerce companies.

1.5. Summary of Chapter 1

Main project goal is to build polarity classification of Russian language' model (as accurate as possible in comparison to state-of-the art performance baselines), which allows assessing users' attitude towards researched Russian E-Commerce companies by analyzing UGC towards company on chosen SNS.. The reason behind choosing this goal was originally based on discovered research gap — the lack of empirical studies upon applications of polarity classification of Russian language for E-Commerce companies. The unique and innovative part of our empirical research is in comparative study of polarity classification of Russian language' models and ongoing step-by-step creation of country-specific and tailored to researched companies' specifics polarity classification model, trained on massive amount of unstructured textual data on CA.

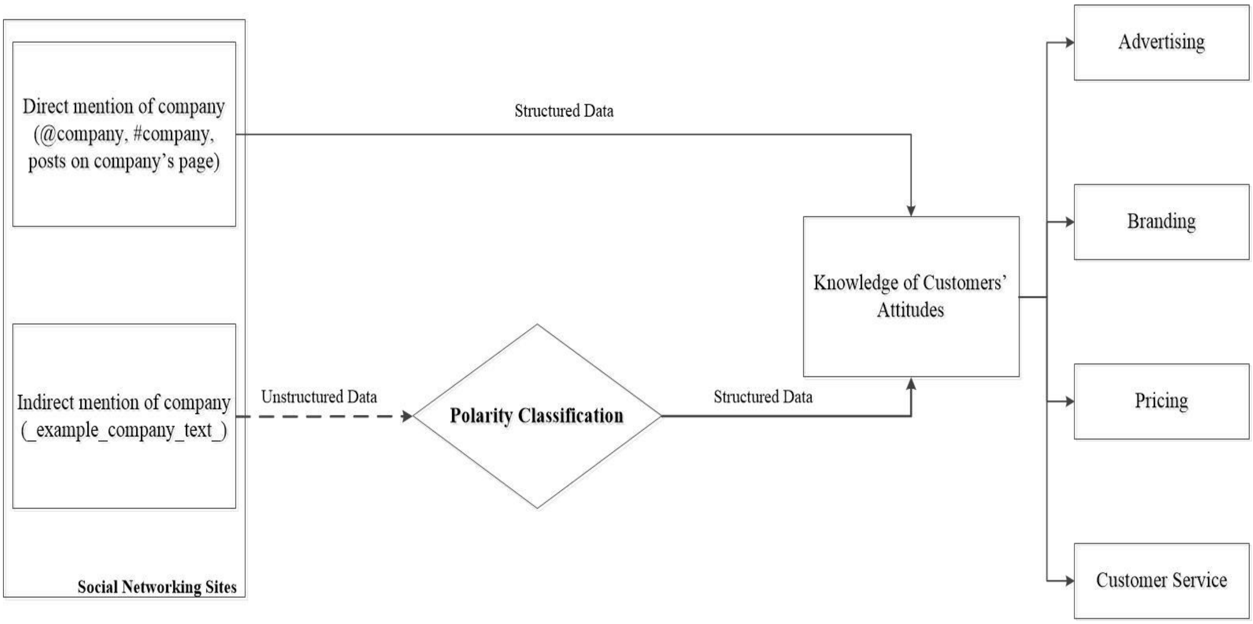

Direct mention of company (@company, #company, posts on company's page)
Indirect mention of company (_example_company_text_)
Social Networking Sites
Structured Data
Unstructured Data
Polarity Classification
Structured Data
Knowledge of Customers' Attitudes
Advertising
Branding
Pricing
Customer Service

Figure 8. Theoretical framework of master thesis

To meet stated research objectives, theory on Customer Attitude and how it affects consumers' desire to purchase specific E-Commerce products and services, along with is ongoing influence on companies' financial and product KPIs was researched. Secondly, Big Data and Data Mining as efficient approaches to CA extraction and analysis were reviewed. This part included overview of Big Data analytics, including its distinctive characteristics, stages of data mining and data analysis processes as well as common tools and techniques for extraction BD from SNS. Finally, Natural Language Processing as efficient tool to work with unstructured data and Sentiment Analysis as a part of NLP specialized in work with textual opinionated data were analyzed. Topic modeling and polarity classification as SA tasks aimed on extraction of aspects and their sentiment evaluation were reviewed. Value of SA for commercial companies was overviewed.

CHAPTER 2. RESEARCH METHODOLOGY

Second chapter is concentrated on creation of research strategy. This chapter consists of 5 key parts — research strategy design itself, business understanding, data understanding along with the process of data preparation, modelling stage and evaluation stage. Through obtaining business and data understanding, business goals, input and output are formulated. During the process of business understanding, RQ1 upon relevant and useful for managers' data output is answered. During data understanding stage, RQ2 upon suitable sources of relevant UGC is answered. On the stage of modelling, existing state-of-the-art polarity classification' models (with usage of lexicon and without it) are reviewed and compared in terms of their performance efficiency to answer RQ3. Suitable for research goal accuracy measurement is identified in order to answer RQ4. In addition, criteria to choose suitable for research goal model are presented in order to answer RQ5. Finally, all the data is prepared and model is chosen to answer RQ6 and RQ7 in empirical part of research in chapter 3. All research questions to be answered are presented on Figure 9.

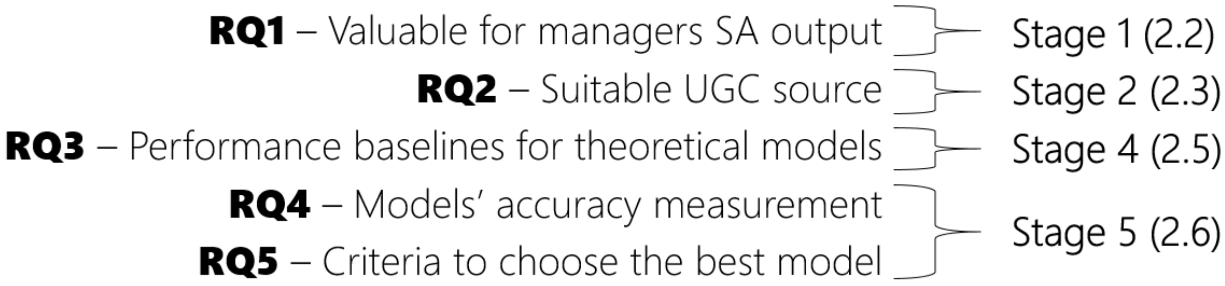

RQ1 – Valuable for managers SA output
RQ2 – Suitable UGC source
RQ3 – Performance baselines for theoretical models
RQ4 – Models' accuracy measurement
RQ5 – Criteria to choose the best model
Stage 1 (2.2)
Stage 2 (2.3)
Stage 4 (2.5)
Stage 5 (2.6)

Figure 9. Research Questions to be answered in chapter 2

2.1. Research Design

Final research goal is to create polarity classification model, which helps managers of E-Commerce companies to get additional knowledge upon Customers' Attitudes towards their brands, products and services. While performing this goal, the research gap settled before is fulfilled. To be able to fulfill research questions, comprehensive understanding of researched E-Commerce companies, Russian segment of SNS, empirical studies upon state-of-the-art polarity classification models and approaches to its accuracy measurement is obtained.

Empirical research design is based upon CRISP-DM data mining methodology and consists of six main stages (Chapman et al., 1999):

1. Business understanding
   a. Determine business objectives — analyze company and formulate business goal

   b. Determine data mining goals — shape desired and reachable input/output

2. Data understanding
   a. Collect initial data — pick suitable source beforehand
   b. Describe data – do descriptive statistics
   c. Explore data – identify needed data points
   d. Verify data quality — prove if data and source are proper

3. Data preparation
   a. Select necessary data
   b. Clean data
   c. Construct data
   d. Format data

4. Modelling
   a. Select modeling technique — compare state-of-the-art models
   b. Generate test design – state inputs and outputs
   c. Build model – build based on theoretical and data review initial models
   d. Assess model — compare with each other and with existing services

5. Evaluation
   a. Evaluate results — choose right accuracy measurements
   b. Review process – return to previous steps if needed

6. Deployment
   a. Deploy model
   b. Produce final conclusions

Steps and goals for all the stages are briefly described at first. Then, all the stages are explained more thoroughly, with the first three stages (business understanding, data understanding and data preparation) being completely done in chapter 2 and last three stages (modeling, evaluation and deployment) being iteratively performed during empirical research in chapter 3. Empirical research design is schematically presented in Figure 10.

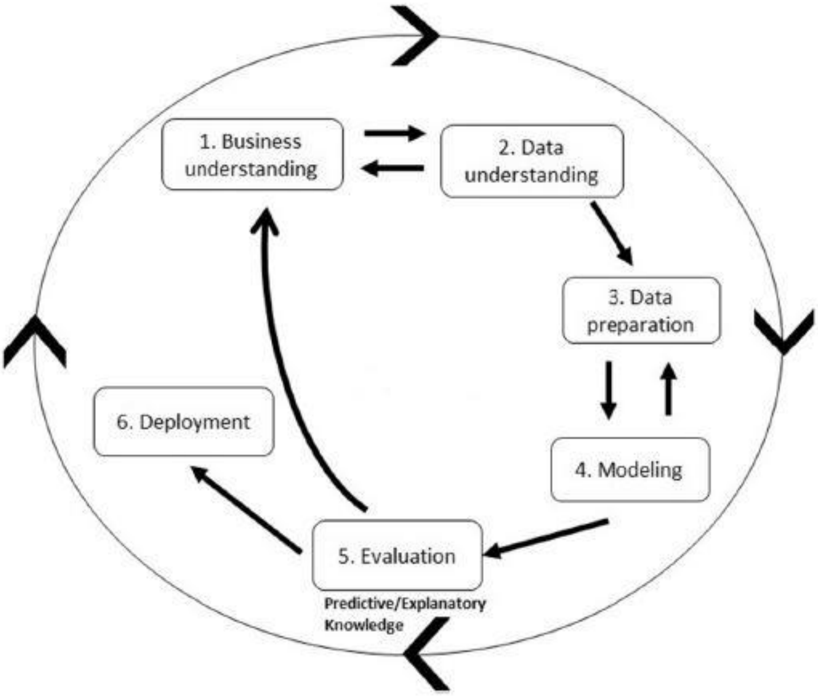


Figure 10. Research design structure
(Source: Chapman et al., 1999)

2.1.1. Business understanding

The first step in data mining process is to obtain business understanding. Goal of this stage is to overview researched companies (including products and services they provide, competitors' activities, their social media presence and digital marketing activities). In addition, it is important to identify specifics of Russian E-Commerce, which may influence analysis of data, accuracy of the model or further business applications. In order to fulfill this stage exploratory type of research (with elements of both qualitative and quantitative exploration) is used. Method — analysis of secondary public data, such as results of market and competitors' analysis, official information from companies' websites, industry experts opinions and data from companies' forums.

The next crucial step is to determine data mining goals. First, to have an initial look at amount and type of input data needed to fulfill our research questions. Due to the specifics of stated research questions, we will work with textual data and focus on its semantics (words, digits, emoticons) and sentiments (attitude / emotional vector behind semantics). In general, massive amounts of user-generated textual data referring to chosen step earlier companies is needed. In addition, this data should be Time Series (historical data will be needed to train machine learning SA models) and include opinionated data (since we are interested in sentiment, there is no need in factual data). There are several criteria for selection of data sources. First, a data source should include respective to abovementioned requirements data. Second, it should be possible for a third party to get an access and pull needed data (e.g. existence of any API or openness to crawlers is critical) in convertible format (.json or .csv or .xml). Finally, we defined the desired final output of the thesis as a verified polarity classification model, which allows assessing users' attitude towards specific topics and is capable to produce output in format 'Topic — Sentiment'.

2.1.2. Data understanding

On the stage of data collection, it is necessary to obtain access to desired data at the first place. Then, since there is already initial data understanding, next step is to design the database (create SQL / NoSQL database where extracted data will be warehoused). Finally, to build a pipeline to extract and save the data to DB.

There will be two types of data needed to be collected — training and testing data. Each of them is initially described (amount of data points, # of unique variables, extractable features) and explored (search for the most suitable data for research purposes). Then, verification of data quality (with regard to data requirements) is performed.

2.1.3. Data preparation

Third stage — data preparation — is focused on selection and preprocessing of relevant for research purposes data. The stage of data preprocessing may be separated into three steps — standard handling, general text preprocessing and source-specific handling procedures. In this case, it includes one additional unconventional step — subjectivity classification procedure.

Standard handling includes:

- Looking for non-relevant data — manual review to identify extra filter criteria, translate non-Russian elements, delete company-shared data;
- Preprocessing dataset — check for missing values, delete duplicates, manipulate labels and transform data to needed formats for further processing.

General text preprocessing includes (not all of these procedures are obligatory / unescapable):

- Tokenization — split sentences and words in the body of text;
- Stop-word removal — filter out useless (in terms of sentiment classification) words;
- Stemming — normalize words via stems extraction;
- Part-of-Speech tagging — label words by part of speech;
- Chunking — group nouns with the related words;
- Chinking — manipulate chunks;
- Named Entity Recognition (NER) — pull entities within text automatically;
- Lemmatization — find another word forms.

Source-specific feature selection and handling procedures are optional (decision upon their application depends on research complexity) and include:

- Removal of urls;
- Removal of user mentions;
- Substitution/detection of hashtags;
- Substitution/detection of emoticons;
- Spelling correction;
- Abbreviation lookup;
- Elongation normalization;
- Punctuation removal;
- Detection of amplifiers and diminishers;
- Negation scope detection and handling;
- Capitalization handling.

Appropriate NLP tools, such as POS and NER taggers, syntactic parsers, lemmatizers, tokenizers, etc., to be used for that. The scope of applicable procedures is defined by modelling type. Subjectivity classification procedure is described in paragraph 2.4. As a main result of these operations, data for each part of modelling stage would be cleaned and ready-to-be-modeled.

2.1.4. Modelling

Since our goal is to create maximally accurate model, this part starts with detailed review of SA models — outcomes of various empirical researches in Russian and English languages are analyzed and described models' performance levels are compared. In addition, in the empirical part, performance levels of existing SA services are analyzed and compared with researched models. After comparative analysis of various models, the criteria to choose the most suitable for research purposes model are formulated.

2.1.5. Evaluation

During the stage of evaluation, research question RQ7 is answered. Practically speaking, first step on this stage is creation of a framework to check accuracy and robustness of results produced by models — use manual verification for dictionary-based SA model and cross-validation (k-fold or Monte-Carlo) for ML and DL models. Then, compare them to some existing services in terms of its accuracy (with Precision, Recall and F1-score as metrics). For final model only, feedback loop with 'Modeling' stage is built, which allows iteratively improving hybrid model to the needed performance level or industry specifics.

2.1.6. Deployment

Deploy the most efficient model and analyze collected data with its help. Deployed model has several managerial applications — provide companies with initial insights upon how users interact with brand online (e.g., the most mentioned/discussed topics related to companies), give insights on users' attitude (e.g., topics and their polarity). It is also possible to extend the model's functionality and perform quantitative analysis of Opinion data (geographies, spatial data, user' characteristics).

2.2. Business understanding

For research purposes, five Russian E-Commerce companies (Wildberries, Citilink, M.Video, Lamoda, Sportmaster) were picked. All of them are ranked within Top-20 of largest Russian E-Commerce companies. Quick overview of these companies:

1. LLC Wildberries is the largest E-Commerce company in Russia. It specialized in fashion clothes — apparel, shoes and accessories — and works with numerous world-known brands (Nike, Boss, Topshop, etc.). It has 1154 self-discharge points in more than 100 cities across Russia. Its sales volume equals 63.8 bln rubles with Average Order Value (AOV) of 1.600 rubles in 2017 (Data Insight, 2017). It has online shop and is presented at such SNS, as vk.com (community with 405.000 subscribers), twitter.com (page with 4.300 followers), Instagram (page with 122.000 followers) and facebook.com (page with 84.000 subscribers). One of specific features of the shop is that it allows its suppliers to sell clothes directly via personal account on wildberries.ru.

2. LLC Citilink is 2nd largest E-Commerce company in Russia. It specializes in electronic appliances — laptops, PCs, smartphones — and works with numerous world-known brands (Apple, Samsung, Microsoft, etc.). It has 50 points-of-sales and 470 self-discharge points in 320 cities across Russia. Its sales volume equals 55.2 bln rubles with AOV of 10.620 rubles in 2017 (Data Insight, 2017). It has corporate website, online shop and is presented at such SNS, as vk.com (community with 169.000 subscribers), Instagram (page with 3.200 followers) and facebook.com (page with 10.000 subscribers).

3. PJSC M.Video is 4th largest E-Commerce company is Russia. It specializes in electronic appliances — laptops, PCs, smartphones — and works with numerous world-known brands (Apple, Samsung, Microsoft, etc.). It has 424 points-of-sales in 169 cities across Russia. Its sales volume equals 36.7 bln rubles with AOV of 10.280 rubles in 2017 (Data Insight, 2017). It has corporate website, online shop and is presented at such SNS, as vk.com (community with 238.000 subscribers), twitter.com (page with 20.000 followers) and facebook.com (page with 85.000 subscribers). Current digital marketing activities include interaction with customers via SNS,

targeted advertising (in VK, YouTube) and context advertising. Recently, M.Video strengthened it positions through acquiring Eldorado (5th largest E-Commerce company in Russia before acquisition). Currently, there is one de-jure owner of those two companies, but companies decided to keep their brands and planned to have separate marketing and sales teams (Khabibrakhimov, 2018). Company's representatives mention that M.Video is oriented on further development of IT-solutions, which will allow company growing online sales' volume (AITC, 2017).

4. Lamoda is 6th largest E-Commerce company in Russia. It is part of Global Fashion Group. It specializes in fashion clothes — apparel, shoes and accessories — and works with numerous world-known brands (Mango, Adidas, Gucci, etc.). It has about 500 self-discharge points in more than 150 cities across Russia. Its sales volume equals 23.6 bln rubles with AOV of 5.860 rubles in 2017 (Data Insight, 2017). It has online shop and is presented at such SNS, as vk.com (community with 307.000 subscribers), twitter.com (page with 8.700 followers), Instagram (page with 231.000 followers) and facebook.com (page with 65.000 subscribers).

5. Sportmaster is 17th largest E-Commerce company in Russia. It specializes in sport-related equipment and apparel. It has more than 450 points-of-sales in more than 200 cities across Russia. Its sales volume equals 10.3 bln rubles with AOV of 7.990 rubles in 2017 (Data Insight, 2017). It has online shop and is presented at such SNS, as vk.com (community with 393.000 subscribers), twitter.com (page with 2.000 followers), Instagram (page with 112.000 followers) and facebook.com (page with 55.000 subscribers).

There are several reasons why these companies were chosen. First, they are in Top-20 of Russian E-Commerce in terms of sales volume and specialized on electronic appliances (Data Insight, 2017). Second, all selected companies specialize on different types of products — apparel, sport equipment, electronic appliances. Third, selected companies have more subscribers within their online communities in SNS, as well as those subscribers act more actively in comparison to other companies within Top-20 biggest Russian E-Commerce companies.

In addition, it is important to clarify that the object of our research are not exclusively mentions of researched companies itself, but also brands or products sold via companies' platforms. This nuance is caused by the reason that E-Commerce companies tangible (and intangible) metrics are built upon products they sell. For example, if price for iPhone within M.Video is unreasonably higher than in Ulmart, customers will not be happy with this and both tangible (sales volume, revenue volume) and intangible (brand comprehension, customer satisfaction level) metrics will worsen.

Thus, initial business goal of empirical part of current research is set up — to get additional knowledge of Customers' Attitudes towards various aspects of selected E-Commerce companies' commerce activities (delivery, service, prices, quality, website, etc.) to make them more efficient. Moreover, additional possible actions may be based on results of current research: this one-time customers' attitude mining may be automated, pipelined and integrated into process of marketing campaigns planning and tuning; for quick analysis performance, an independent service or platform may be built, deployed and used within different departments.

2.3. Data understanding

First step of this stage was to choose appropriate data source. Resources to get data from — online reviews from company's online shop, VK, Twitter, Facebook. All these sources fit data selection criteria (acceptable and allow storing in proper format). Online reviews contain a lot of info purely about brands, though it may be a good extra source (due to its large volumes). VK does not fit due to specifics of VK communities and mechanics of interaction with UGC — M.Video community moderators see all the company-related UGC and can work with it within one page. Facebook is more a one-way (company-to-customer) interaction channel and amount of sentiment bearing content is close to zero.

Twitter suits both data and source selection criteria. SA of data from Twitter is one of the topics, which is especially thoroughly researched in last years (Barbosa & Feng, 2010; Bifet et al., 2011; Kouloumpis et al., 2011; O'Connor et al., 2010). Twitter is a microblog platform with 328 million monthly active users worldwide (8 million monthly active users in Russia) and specifics, such as limitations in number of characters (140 characters max) within message (Statista, 2018). Due to specifics of chosen SNS, UGC contains a lot of spelling mistakes, acronyms, targets (@) and hashtags.

There are two major ways users on Twitter interact with researched companies — via targeted communication (text has format '@M.Video __example-text_ _' or '#мвидео__example-text__'), or via indirect mentions (text has format '__example-text-мвидео-text'). While company representatives tend to react and handle targeted interaction, it completely does not respond to indirect mentions (0 replies for last 100 indirect mentions). This indicates a gap in knowledge about valuable customer attitude (its value is high due to the facts that Russians on Twitter tend to be radically honest about their opinions and emotions).

After data source is selected, it is vital to understand what types of input and output are needed. All the input data consists of two parts — training data, which consists of lexicon (lexicon is the same thing as dictionary) for dictionary-based models, labeled text corpora for Subjectivity

Classification, labeled tweet corpora for supervised ML SA models and unlabeled text corpora for semi-supervised ML SA models, and testing data that is relevant to topic, real world and similar for all models (to avoid biases during evaluation and comparison).

Several preparations for the stage of data collection step were made:

- Twitter API' keys and credentials were obtained – see (Appendix 1. Getting access to Twitter data);
- Data formats were chosen and databases were designed – see (Appendix 2. NoSQL Cloudant database view);
- Process of data collection was pipelined – see (Appendix 3. Node-Red data collection).

Training data was collected at the first place.

Lexicon for dictionary-based SA models (Russian words' corpus):

While there are several comprehensive wide-purpose Russian language lexicons (Yablonsky, 2003), there is still lack of Russian language domain-specific lexicons made specifically for SNS analysis with the most of lexicons being in closed access. Two different lexicons (seed and extended) were collected to be able to benchmark different corpora and avoid data overfitting in case of extended word corpus.

Seed lexicon was comprised of one general-topic lexicon of labeled by experts' emotional Russian language — Linis-Crowd (≈ 27.000 terms; polarity weights from -2 for strongly negative to +2 for strongly positive). Example of seed lexicon in (Appendix 4. Raw seed lexicon' example).

Extended word lexicon was comprised of abovementioned seed lexicon, then enriched with domain-specific lexicon (#Russian #E-Commerce #electronics), collected at the stage of data understanding, and in addition enriched by general-purpose Russian Sentiment Lexicon (RuSentiLex, 2017). There are numerous options aimed on enhancement of lexicon, though most of them are time-consuming — for example, it can be enriched by narrow-topic lexicon from Russian Wikipedia latest dump or by translated from English lexicons (e.g., 'wordnet'). Example of extended lexicon in (Appendix 5. Raw extended lexicon' example).

Data for supervised Machine Learning SA models (marked up texts):

For supervised ML SA models, text corpus with sentiment markup (from Linis-Crowd — do not confuse with identical name word lexicon used before) and general-purpose tweet corpus from SentiStrength collection with sentiment markup in amount of 19300 labeled texts were collected. For future researches, it is preferable to do sanity check with industry expert. Example of data for supervised training in (Appendix 6. Raw labeled text corpora).

Data for unsupervised Machine Learning SA models (unlabeled tweets):

For unsupervised DL SA models, unlabeled text corpus was needed. It was possible to get this sample through one-time data request (2017-2018) via Twitter Historical PowerTrack API. However, request was rejected. Next attempt was to grab data through Twitter FireHose with trial access to 'sentimentmetrics.com'. This request was rejected as well. It was decided to pick up some ready-to-go corpora. Text corpora of unlabeled tweets in .txt format and in amount of 340000 rows was collected from MorphoRuEval competition (MorphoRuEval, 2017). Example of unsupervised training data in (Appendix 7. Raw unlabeled text corpora).

Test tweet samples for researched companies:

Real world relevant testing data was collected for the period of 1 January — 30 April 2018. Initially, it was planned to collect testing set through real-time data gathering via Twitter Streaming API in standard format .json. Since amount of companies increased after some changes in research process, test sample was collected with the help of Python script and library Twitter (PyPi, 2018). Test sample was collected in amount of 5143 tweets with indirect mentions of all researched companies (Appendix 8. Code for Testing tweets dataset collection). Initial raw sample included both subjective and objective data. Since data was collected from non-related to companies' accounts pages, it represents a true sample of actual tweets in terms of language use and content (Agarwal, 2011). (Appendix 9. Raw testing tweet sample)

To sum up, Table 1 contains different types of input and output that are used for this research.

Table 1. Input/output data for polarity classification models

Part of data	Input	Output
Seed lexicon	Unstructured text corpora	Labeled text corpora
Extended lexicon	Unstructured text corpora + domain-specific and SNS corpora	Labeled text corpora
Subjectivity Classification	Unstructured textual data (UGC) from SNS	Pool of opinionated (subjective) tweets
Lexicon-based models	Pool of opinionated tweets + Seed & extended lexicons + Rules' set	Topic — Sentiment + Model Accuracy

Table 1 (continued). Input/output data for polarity classification models

Part of data	Input	Output
ML-based models	Pool of opinionated (subjective) tweets + Labeled text corpora	Topic — Sentiment + Model Accuracy
DL-based models	Pool of opinionated (subjective) tweets + Unlabeled text corpora	Topic — Sentiment + Model Accuracy

Finally, all the texts were collected and transformed to convenient for further work format of .csv and .xlsx documents.

2.4. Data preparation

The key point behind selection process was in interconnection of data preparation stage with modelling stage. Only after getting understanding about the fact of which model is to be built, it is reasonable to actually start selecting and preprocessing data. However, it is quite common situation, when desired model cannot be created due to the lack of necessary prepared data or its insufficient quality. That is the reason, why all possibly needed data was collected and prepared.

For preprocessing tasks, Python programming language was used. In specifics, Python Pandas library was used to standard handling, while for general text processing and source-specific processing Python NLTK library (with such popular corpora as 'brown' and 'wordnet' initially installed) was used. NLTK library includes such tools, as tokenizers, POS and NER taggers, lemmatizers and stop-words collections, which makes it perfect for text processing.

Lexicon for dictionary-based SA models (Russian words' corpora):

First step was to prepare seed lexicon through preprocessing raw Linis-Crowd lexicon. To perform standard handling were used following methods: duplicate deletion, label manipulation, transformation to DataFrame type. Since initial polarity was 5-way (-2 to +2), it was reduced to 2-way polarity (0 to 1). Since we needed only word-sentiment pairs, no further data processing was required. Example of prepared seed lexicon in (Appendix 10. Prepared seed lexicon' description).

Second step was to prepare extended lexicon through enriching prepared Linis-Crowd lexicon with domain-specific lexicon and translated English lexicon. Generation of domain-specific lexicon was partially performed using semi-automatic methods (e.g. bootstrapping). Translation of English lexicon was performed with the help of Google Translate bulk queries. To both domain-specific and translated lexicons the same standard handling procedures were applied. Example of prepared extended lexicon in (Appendix 11. Prepared extended lexicon' description).

The quality of lexicons was assessed via comparing human text scores with the automatically obtained scores (in case of Linis-Crowd) or via manual verification.

Data for supervised Machine Learning SA models (marked up texts):

First step was to preprocess raw Linis-Crowd text corpus. To perform standard handling were used following methods: duplicate deletion, label manipulation, transformation to DataFrame type. Since this corpus would be used for further vectorization, it was additionally processed with tokenization and stemming. Example of prepared supervised training data in (Appendix 12. Prepared training dataset for supervised learning' description).

Data for unsupervised Machine Learning SA models (unlabeled tweets):

Since during data collection, the preprocessed text corpora were chosen, only standard handling was needed. To perform standard handling were used following methods: duplicate deletion, label manipulation, transformation to DataFrame type. Example of prepared unsupervised training data in (Appendix 13. Prepared training dataset for unsupervised learning' description).

Test tweet samples for researched companies:

Duplicated are removed. Redundant whitespaces are deleted. Text is lowercased. Punctuation was removed (including exclamation signs, because it goes out of the research scope). All Latin symbols are translated (e.g., '#M.Video' translates to 'мвидео').

Text was tokenized. Stop words were acquired from Python NLTK corpus and partially deleted. Emoticons were deleted (due to problems with Cyrillic encodings — no research handling this issue was found). Negation was handled (tag 'NOT' was added). URLs were removed. Hashtag words are added to tweet with hashtag symbol itself is deleted (Appendix 14. Prepared testing tweet dataset' description).

Data for Subjectivity Classification model:

As it was mentioned in literature review, preliminary subjectivity classification will be performed to split our data to objective and subjective subparts. Subjectivity classification is a tool to extract opinionated tweets out of all gathered textual data. It helps to split data into opinionated and factual.

In perfect situation, to perform SC automatically, large corpus of marked up (2-way markup, 'Subjective' — 'Objective') texts or tweets along with opinion lexicon are needed. This kind of data may be partially collected from Linis-Crowd (along with terms, it also is a collection of marked up Russian texts) and enriched with Russian tweet corpus from conference 'Dialogue-2016'. In sum, training data corpus sufficient to train ML model for subjectivity classification may consist of 35000+ labeled tweets and texts in Russian language.

Subjectivity classification was performed on preprocessed testing tweet sample to get pool of opinionated tweets needed for further research. In this research, due to the scope of research and size of testing sample, it was decided to perform subjectivity classification manually. However, in case of larger datasets to be analyzed or in case manager wants to pipeline the process and minimize amount of work performed by him, subjectivity classification should be made under following framework.

Subjectivity classification is performed at sentence-level. To calibrate algorithm, it is important to remember, that the most appropriate for sentiment analysis is a pool of opinion-bearing rather than simple set of subjective data. That is why after initial 'subjective' — 'objective' dataset' split, additional extraction of opinions (along with opinion aspect and opinion holder) should be extracted.

Subjectivity classification may be implemented with the help of abovementioned corpora and ML models, such as SVM, or be rule-based and apply different patterns for subjectivity identification. For example, (Kravchenko, 2012) built subjectivity classification model based on common patterns within Russian language with additional rule-based filtering and achieved 80% accuracy on dataset, comprised of earphones and cameras' reviews. The main output — 'Holder-Opinion' textual pair — is used in further sentiment classification. The framework for Subjectivity Classification is schematically explained in Figure 11.

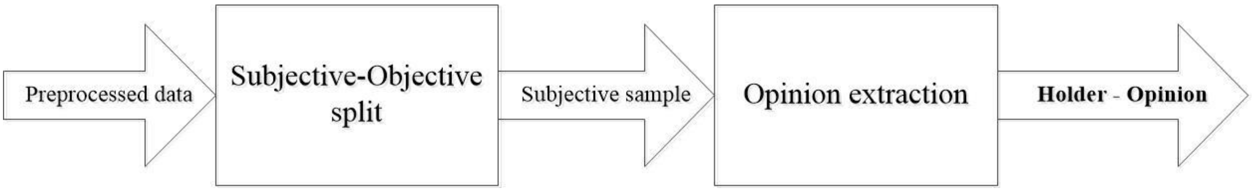


Figure 11. Subjectivity Classification generalized framework

After subjectivity classification was performed, amount of opinionated and factual tweets was calculated, its results may be found in Table 2.

Table 2. Subjectivity classification results

Company	Overall tweets	Subjective	Percentage
Wildberries	580	86	15%
Citilink	118	47	40%
M.Video	695	120	17%
Lamoda	3126	227	7%
Sportmaster	624	154	25%

Large percent of tweets, made by Lamoda customers, contain only objective information, related to coupons, sales announcements and company-related news (new acquisitions or CEO-related info). In addition to amount of collected subjective data (even on monthly basis), most of it is direct mentions of Lamoda handled by its customer service specialists. Those are reasons why in empirical part only four other companies are analyzed.

2.5. Modelling

On this stage, initial action is to select modeling technique for polarity classification of Russian language. For every approach (lexicon-based, ML-based and DL-based), current research upon every model and its components are reviewed. Different methods to calculate model performance and respective performance baselines are analyzed. In addition, due to specifics of desired output, topic modeling procedure is performed at the beginning of modeling step to extract opinionated topics from dataset. Finally, test design specifically for models with best performance is generated. This design was created with consideration of research specifics (Twitter as main data source; flexible and morphologically tricky Russian language).

2.5.1. Topic modeling

Classifying text at the document level or at the sentence level does not provide the necessary detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; to fulfill this task aspect level analysis is needed. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities (Medhat et al., 2014). The key idea of topic modeling is to extract important aspects out of sentences (or clauses of sentences). This extraction procedure may be based on different criteria — aspect selection based on POS tags, n-grams, or other features, represented in the form of bag-of-words matrix. This is why topic modeling process may be separated into two steps — collection of bag-of-words matrix (with the help of such tools as TF-IDF transformers) and topic modeling itself (with preselected topics and its number as variables — or without them). Because of topic modeling, several clusters of topics are extracted.

There are three specifics to the task of topic modelling in this research: there is no a priori knowledge upon aspects to be extracted, which makes it possible to use the same technique as applied by (Farhadloo et al., 2016); there is no need in calculation of aspect weights (the final output just to be in format ‘Topic — Sentiment’); there is no need in identification of tweet author. These are three reasons why Latent Dirichlet Allocation (LDA) is used as topic modelling algorithm. It is unsupervised probabilistic model and its main idea is to cluster different words on the basis of their sparsity in the analyzed sentences. Method to assess accuracy of topic modeling in current research is straightforward manual verification (and sanity check).

2.5.2. Polarity classification modeling

In general, there are several approaches to SA problems — with usage of lexicon and without it. Approach with usage of lexicon includes rule-based and dictionary-based models. Approach with not using lexicon rely on ML and DL techniques and may be split into supervised, semi-supervised and unsupervised. They are compared in Table 3:

Table 3. Comparison of DM modelling methods

Approach	Classification	Pros	Cons
Rule-based	Supervised and unsupervised	+No need in classifier creation and training	-Accuracy depends strongly on rules -Noisy SNS data worsen model’s accuracy
Dictionary-based	Unsupervised	+No need in classifier creation and training	-Requires huge marked up text corpora -For competitive accuracy classifier still needed
ML/DL-based	Supervised, semi-supervised and unsupervised	+No lexicon needed +High accuracy	-Domain-specific -Need to retrain for another sphere

The fundamentals, basic terminology and important steps (such as feature selection, weight assignment and classifier selection and creation) needed to understand polarity classification process were reviewed before in chapter 1. Before initializing the process of models’ performance analysis, the processes for every type of models are briefly described. After that, performance levels of state-of-the-art models are analyzed. Criteria for choice of those models are presented in the beginning of each performance analysis section.

Lexicon-based SA models' description

First, dictionary-based models are described. This type of models usually are applied on sentence-level, not getting more granular. The key idea is to identify existing words (other POSs) and calculate their overall sentiment within sentence, which is the final output of model ('Sentence — Sentiment'). Creation of dictionary-based model with different lexicons and unstructured textual test sample as key inputs is described in Figure 12 and includes next steps:

- Lexicon generation, where training corpus is gathered. This corpus may consists of already existing corpora (every element of it must be marked up with some sentiment) or some lexicon pieces are specifically created and labeled with some sentiment. If vectored corpus is used, then features are selected on this stage.
- Testing data collection and preparation. Selected features within testing data should correspond to lexicon features (if lexicon consists of labeled n-grams, then testing data should consists of to-be-labeled n-grams).
- Final Evaluation. Testing data is evaluated, overall sentiments counted, accuracy results calculated.

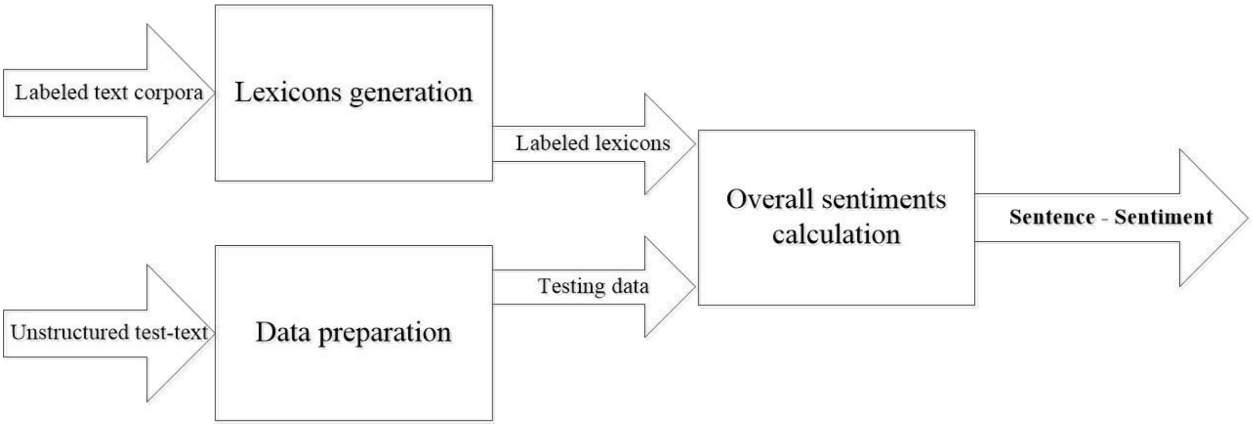


Figure 12. Dictionary-based SA model framework

Second, rule-based models are described. This type of models usually are applied on sentence-level, not getting more granular. Rule-based approach is advanced dictionary-based approach. The key idea is to calculate overall sentiment within sentence, with regard to rules upon sentiment evaluation and labeled lexicon. Creation of rule-based model with different lexicons, rules upon sentiment evaluation and unstructured test sample as key inputs is described in Figure 13 and includes next steps:

- Lexicon generation, where training corpus and set of rules are gathered. Rules are usually applied on the stage of testing data preparation (some words are deleted, some tags are added), but some rules correspond to the stage of final evaluation (several specific words in a row may lead to a different final sentiment).

- Testing data preparation. The same as for dictionary-based models, but with regard to rules' requirements.
- Initial evaluation. Words within sentences are extracted and labeled accordingly to exploited lexicon.
- Final evaluation. The same as for dictionary-based models, but with regard to rules' requirements. In some cases, to apply rules to testing data, classifiers are trained (still, classifier are not necessary for rule-based model to function).

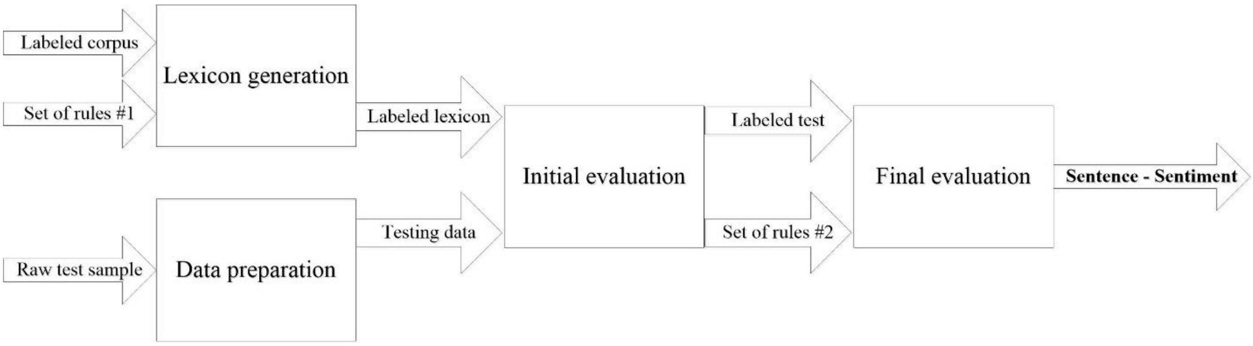


Figure 13. Rule-based SA model framework

Lexicon-based SA models' performance analysis

Overview of lexicon-based models' performance in current research includes analysis of state-of-the-art research towards dictionary-based and rule-based models. Performance of both rule-based and dictionary-based polarity classification models depends on quality and relevance of lexicon in case of dictionary-based models and set of rules along with ML classifier in case of rule-based models. Since the overall accuracy of dictionary-based models depends heavily on lexicon relevance in relation to task performed, lexicon generation is relevant in scope of this work as long as it is related to Russian language or is applied to analysis of Twitter data. As it was mentioned before, only the most accurate models are analyzed.

There are numerous researches upon collection and creation of general and domain-specific lexicons along with evaluation of its accuracy in both Russian and English languages. Numerous studies focused on creation of lexicons may be found. (Saif et al., 2016) proposed a model capable of updating words' sentiments based on their context (named SentiCircle) and tested it on Stanford Twitter Set (STS). The final setup consisted of SentiCircle + SentiWordNet lexicon. Macro F-measure of 85,45% for STS was achieved. (Keshavarz & Abadeh, 2017) described state-of-the-art approach to lexicon generation (named 'adaptive lexicon learning by genetic algorithm (ALGA)') and tested it on several popular pre-labeled datasets in English language, including STS and

Sanders (collected on #apple, #microsoft and #google tweets). The final setup consisted of ALGA + n-gram features. Macro F-measures of 84,24% for STS and 85,2% for Sanders were achieved. (Deng et al., 2017) suggested model based on method, which allows adapting existing sentiment lexicons for domain- or language-specific sentiment classification. This model was tested on corpus of tweets mentioning S&P 500 companies on Twitter. The final setup consisted of Combined4_ARVN lexicon and final result was calculated via straightforward lexicon scoring approach. Macro F-measure of 80,31% for business tweet set was achieved. (Balahur & Perea-Ortega, 2015) created multilingual dictionary-based polarity classification model and tested it on English and Spanish tweets. Several approaches to training data preparation were compared (slang replacements, punctuation mark-up, usage of domain dictionaries). The final setup consisted of bigram features and SVM as classifier with no slang replacements and usage of domain lexicon. Macro F-measure of 69% was achieved.

Rule-based models are less popular and in many cases are more complex. (Karpov et al., 2016) used set of rules only for data preparation stage and did not use any classifier on the stage of final evaluation. Model testing was performed on Russian language text corpora upon banks and telecom industry' reviews. The final setup consisted of word2vec-based features + set of domain rules and final result was calculated via straightforward lexicon scoring approach. Macro F-measures of 49% for telecom industry tweets and 45,9% for banking industry tweets were achieved. Sudden decrease in accuracy (comparing to dictionary-based methods) shows low level of development of feature selection tools and dictionaries for Russian language' richness. For polarity classification of the same telecom industry tweets, (Vasilyev et al., 2016) built model upon fragments rules' set. During the test, different ML classifiers were compared in terms of accuracy. The most accurate setup consisted from fragment rules model + SVM classifier. Macro F-measure of 35,3% was achieved.

Machine Learning and Deep Learning SA models' description

This type of models may be applied on sentence-level, but to get more precise results, it is recommended to apply more fine-granular target, such as clauses of sentence. The main differentiator of these models is the availability of 'teacher' — (automatically or manually) marked up training data. While supervised learning require the availability of a huge labeled relevant to domain corpora, unsupervised learning is working with smaller and general training corpora. First, supervised ML and DL models are described. The key idea is to build vector representation of words from the test sample, assign weights and train classifier on labeled training datasets (they may be general, but usually the more domain-specific they are — higher the final accuracy is). The main drawback of these models is in necessity of human efforts to label the training data by assigning proper weights to thousands (or even millions) of words. Creation of ML-based model with labeled training data and opinionated test sample as key inputs is described in Figure 14 and includes next steps:

- Training data preparation. Selected features within testing data should correspond to lexicon features (if lexicon consists of labeled n-grams, then testing data should consists of to-be-labeled n-grams);
- Testing data preparation. The same as for dictionary-based models;
- Feature selection. It may be tedious to perform feature selection from scratch. However, there are several prepared labeled vector collections, such as Russian Distributional Thesaurus and RusVectores (Rusvectores, 2018) word embedding' libraries, as well as tools to translate them into proper training data, such as word2vec and doc2vec;
- Weight assignment. Since supervised learning use labeled training data, to each vector corresponding value is assigned;
- Classifier choice and training. It is the same as in ML NLP — probabilistic classifiers (Naïve Bayes, Maximum Entropy) or linear classifiers (SVM, Linear Classifier) may be used;
- Initial evaluation. Initial model testing and scoring;
- Model improvement. Through iterations and changes for every step of modeling process (feature selection, weight assignment, classifier training), the model is made more precise);
- Final evaluation. Testing data is evaluated, accuracy results calculated.

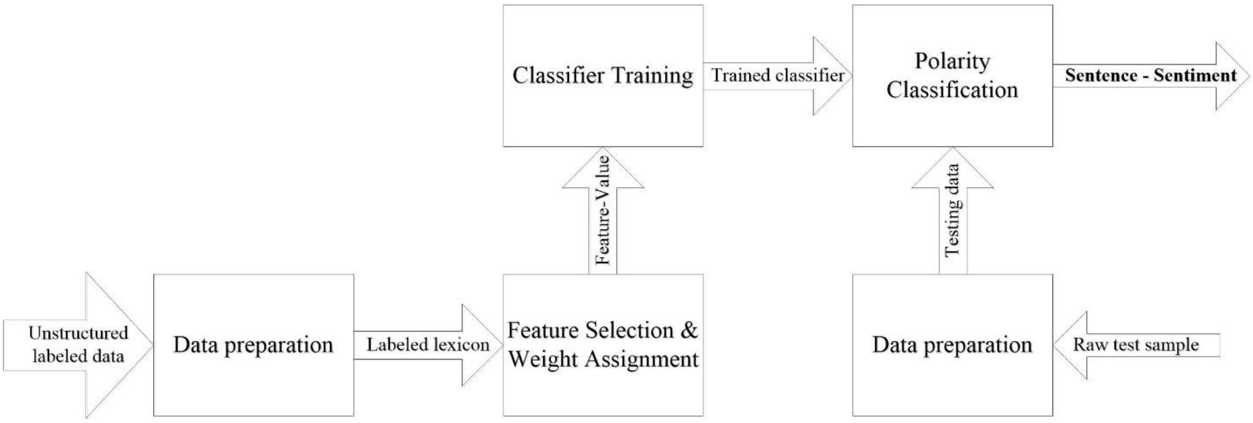


Figure 14. Supervised ML/DL-based SA model framework

Second, unsupervised ML- and DL-based models are described. The main difference between supervised and unsupervised models is in approach to feed the machine with training data. ML-based models use preprocessed and manually verified data with explicit labeling of inputs provided and outputs related to those inputs. If it is classification task, then all the inputs would relate to one or another class, and the output is strictly framed — it is relation to some class. In case of unsupervised learning, there is no previous knowledge about inputs and outputs, as well as no labeled data. That is why the most suitable for unsupervised text analysis task is clustering. For example, described above word2vec and doc2vec tools are based on this principle. For DL-based models, it is important to choose correct number of epochs, since overtraining lead to worse results during testing. The process of unsupervised ML/DL-based models differs at the early stages, with later stages being quite similar. It is described in Figure 15 and includes next steps:

- Training data preparation. Unlabeled training data, which may be enriched later with features or based on some lexicon corpora values;
- Testing data preparation. The same as for dictionary-based models;
- Feature selection. It may be tedious to perform feature selection from scratch. However, there are several prepared labeled vector collections, such as Russian Distributional Thesaurus and RusVectores (Rusvectores, 2018) word embedding' libraries, as well as tools to translate them into proper training data, such as word2vec and doc2vec;
- Classifier choice and training. The same as in DL NLP classifiers (Recurrent Neural Networks with LSTM or GRU network as a basis; Convolutional Neural Networks) may be used;
- Initial evaluation. Initial model testing and scoring;
- Model improvement. The same as for ML-based models;
- Final evaluation. The same as for ML-based models.

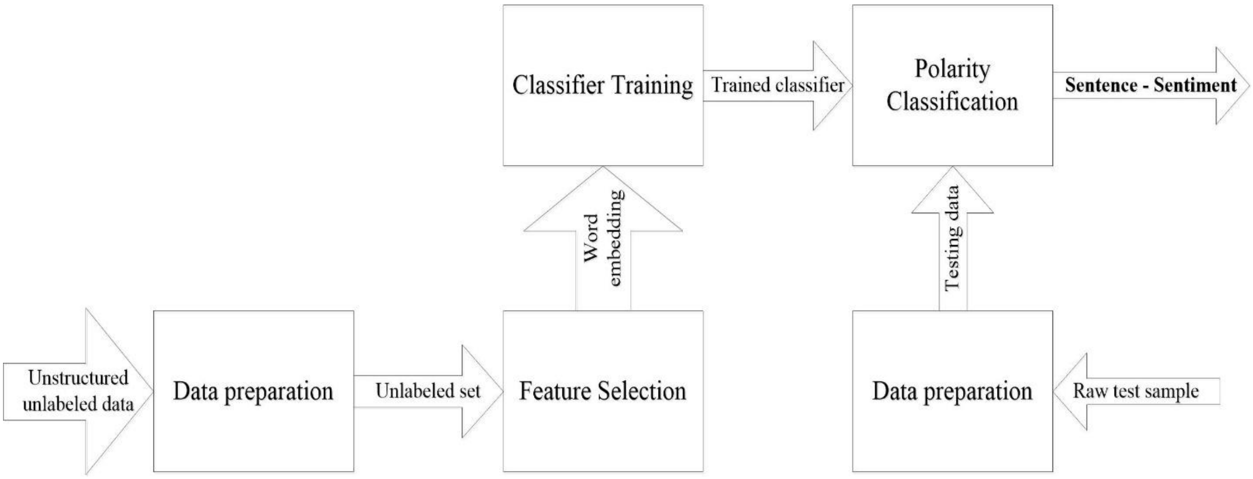


Figure 15. Unsupervised ML/DL-based SA model framework

Machine Learning and Deep Learning SA models' performance analysis

In recent studies, use of ML and DL showed highest results in terms of model accuracy while applied for SA of English language comparing to pure lexicon-based models. Performance of ML and DL based models depends on several factors — quality of training data, efficiency of chosen feature selection method, accuracy of weight assignment and accuracy of classifier. Especially popular in research (in English, Russian and multilingual studies) are ML-based models. (Vilares et al., 2017) created multilingual sentiment analysis model (capable of automatic subjectivity and language detection) and tested it on set of tweets in English and Spanish. Different approaches to bag-of-words feature selection were compared during the experiments — words as features (W), lemmas as features, psychometric properties (P) as features and POS tags (T) as features. The final setup consisted of selected features and linear classifier. Macro F-measure of 69% for combined usage of W, P and T features was achieved. (Rubtsova, 2017) implied ML-based model for analysis of Russian language short texts (online reviews). During research, several methods of feature selection and approaches to collect training data were compared. It was proved that word2vec is the best solution to select features and assign weights in regard to short-text classification tasks. The final setup consisted from word2vec as feature selector and SVM classifier. Macro F-measure of 72% was achieved. (Loukachevitch & Rubtsova, 2016) performed overview of applications of ML-based models used for sentiment analysis on SentiRuEval-2016 competition. Testing data was combined from tweets in Russian language upon telecom and banking industries. One of the best runs showed the model, which consisted from n-grams features + SVM classifier. Macro F-measure of 54,9% for telecom industry and 52,5% for banking industry. Again, even with usage of lexicon-independent ML classifiers, decrease in accuracy between models for English and Russian is undeniable. In addition, such difference is caused by specifics of UGC — tweets with numerous abbreviations, specific slang and misspellings. This is the main reason why only models, applied in current studies for analysis of Twitter messages in Russian language, participate in comparison for the best performance of theoretical models below.

Overview of DL models' performance in current research usually touches upon applications of Forward Neural Networks, Convolutional Neural Networks and Recurrent Neural Networks. (Arkhipenko et al., 2016) studied performance of RNN as classifiers. During research, different configurations of RNN, such as LSTM and GRU were compared. Testing data was combined from tweets in Russian language upon telecom and banking industries, used by many researchers during SentiRuEval-2016. The best run showed GRU network based solution with reversed sequences. Macro F-measure of 55,9% for telecom industry and 55,1% for banking industry were achieved. (Karpov et al., 2016) studied performance of CNN. The final setup of his work consisted of POS-tagged training set and CNN (trained for 40 epochs with stochastic gradient descent). Macro F-measure of 52,6% for telecom industry and 53,6% for banking industry were achieved. From previously mentioned research by (Loukachevitch & Rubtsova, 2016), there is one described DL-based model, which outperformed all other models at SentiRuEval-2016. This model was based upon LSTM variation of RNN classifier and tested on Russian language tweets' datasets. The final setup consisted of words2vec feature (trained on external collection of Russian language tweets and online reviews) and RNN classifier. Macro F-measure of 56% for telecom industry and 55,2% for banking industry were achieved. (Smirnova & Shishkov, 2016) tested RNN as classifier on Russian language tweets' dataset. The final setup consisted of two linked RNNs. Macro F-measure of 69% was achieved.

In addition, researchers tend to use both baseline ML and DL models and their ensembles. Ensemble model denotes hybrid model, which combines elements from both lexicon-based models and ML/DL-based models. Several examples of such approach with Russian language and/or Twitter (short-size) data showed quite high accuracy (not all them, though). In (Arkhipenko et al., 2016) along with basic RNN models, ensemble model of RNN and CNN was created. Macro F-measure of 54% for telecom industry and 53,5% for banking industry were achieved. In addition to RNN deployment, (Smirnova & Shishkov, 2016) also tested ensemble of CNN and RNN with training dataset and testing sample of not related to business pool of tweets. Macro F-measure of 71% was achieved (2% higher than straight RNN model F-measure).

Out of analyzed models, it was necessary to choose the most efficient ones for further empirical research. There were several performance criteria, which were exploited to choose the final model, such as:

1. High F-measure — since it is the main metric for final model evaluation;
2. Suitability for Russian language — due to specifics of Russian language, some models may a priori have less accuracy. As a prove for this fact, during models' overview severe performance gaps between models for English and models for Russian were identified — to avoid these gaps, tested for Russian language models were given higher priority;
3. Suitability for Twitter textual data' specifics — paralinguistic content, large amount of slang, common misspellings to be handled;
4. Availability of training data — in some cases it is not feasible to obtain some sort of necessary data for Russian language — or it is too small to train classifier.

Considering all these criteria, it was decided to pick those models, which showed the highest accuracy while being tested on Russian language tweets. In addition, it should be possible to collect sufficient amount of training data for those models.

To sum up, the best performance results for different types of models, applied to twitter data in Russian language for business-related datasets, are listed in Table 4:

Table 4. Theoretical models' best performance

Approach	Feature Selector	Model classifier	**Macro F-measure**
Rule-based	word2vec	Lexicon scoring	0,49
ML-based	**n-gram**	**SVM**	**0,55**
DL-based	**word2vec**	**RNN**	**0,56**
Ensembles	word2vec	RNN+CNN	0,54

Stages 4 and 5 of data mining process are interconnected and, in the scenario of final model, looped and repeatable. For the ongoing modeling, only opinionated data will be used. Second step on the stage of the modelling is polarity classification itself — it is performed in empirical part of research with ML- and DL-based models only.

2.6. Evaluation

For every modeling approach and modelling step, there are specific metrics on accuracy measurement. However, for every type of models, it is necessary to state number of polarity classes (flat or 2-way — hierarchical or 3-way — ordinal regression or k-way) to be identified. Along with development of NLP, polarity classes move from classification (flat and hierarchical for lexicon-based) to ordinal regression (ML-based). Nevertheless, the choice of the number of polarity classes defines final model accuracy in first place — no matter if it is lexicon-based or not. It is logical that if number of classes is lower — ceteris paribus — the accuracy is higher. This tendency may be observed in numerous researches and competitions, for example, during ROMIP-2012 all the models with 2-way polarity showed higher performance (Chetviorkin & Loukachevitch, 2013). Considering the lack of need in achieving ordinal numerical sentiment result, 2-way polarity scale (‘positive’ — ‘negative’) is used in current research. In addition, key metrics for each type of polarity classification models are described:

Lexicon-based SA models' accuracy

- Accuracy range: 0.00 (min) - 1.00 (max);
- Data (Lexicon and testing) accuracy evaluation procedure: manual verification (human polarity scores vs. automatically obtained scores);
- Overall model accuracy basis: training data (incl. lexicon) accuracy.

Machine Learning and Deep Learning SA models' accuracy

- Accuracy range: 0.00 (min) - 1.00 (max);
- Data (training and testing) Accuracy evaluation procedure: k-fold cross-validation (split on validation and testing set);
- Overall model accuracy: feature selection accuracy + classifier accuracy.

To calculate overall model accuracy, several techniques are used. Since the choice of technique affects final accuracy interpretation, it is important to choose proper one. The choice depends mainly on specifics of testing dataset to be evaluated. If testing sample is balanced (amount of negative and positive pieces of data is roughly equal), then ROC-AUC (Receiver-Operating-Characteristic Area under Curve) is applied. In case of class imbalance (applicable to current research), PR AUC (Precision-Recall Area under Curve) is applied. In addition, this technique suits binary classification tasks and is widely applied for similar SA tasks in NLP and ML communities (Sokolova & Lapalme, 2009).

The final numerical metric of PR AUC is Dice coefficient, a.k.a. F1-score. It helps to identify relations between dataset's positively labels and classifiers' produced labels. To calculate it, values of Precision and Recall are needed. All of them are based on various ratios of True Positive (T_p), True Negative (T_n), False Positive (F_p) and False Negative (F_n) results (together forming 2x2 confusion matrix). In addition, simple Accuracy metric may be useful to get quick understanding of how efficient classifier is. Here is what these metrics show:

- Accuracy shows overall effectiveness of applied classifier. Formula of accuracy in Figure 16:

$$accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

Figure 16. Accuracy calculation formula

- Precision is the measure of result relevancy (the percentage of correctly labeled input data). Precision of T_p is defined as the number of True Positives over the sum of True Positives and False Positives. Formula of precision in Figure 17:

$$precision = \frac{T_p}{T_p + F_p}$$

Figure 17. Precision calculation formula

- Recall is the measure of how many truly relevant results are returned (the percentage of correctly identified labeled input data). Recall of T_p is defined as the number of True Positives, over the sum of True Positives and False Negatives. Formula of recall in Figure 18:

$$recall = \frac{T_p}{T_p + F_n}$$

Figure 18. Recall calculation formula

- The F1-score is the harmonic mean of precision and recall, and is the most common metric for evaluation of ML and DL classifiers. Formula of F-measure in Figure 19:

$$F_1 score = 2 * \frac{precision * recall}{precision + recall}$$

Figure 19. F1-score calculation formula

F1-score may be calculated in different ways. One of them is micro F-measure (calculates metrics for each label and find their unweighted mean) and second is macro F-measure (calculates metrics globally and is applicable in case of 2-way polarity classification formula, which is described above). The choice of F1-score type to be applied depends on various factors — number of labels within classification task (two labels — ‘positive’ and ‘negative’ — in current research), how unbalanced testing data is (data skewness towards ‘negative’ label in current research). In modeling

With the help of these metrics, the accuracies of existing services and created models are calculated and compared during empirical research.

2.7. Summary of chapter 2

Research strategy was designed with accordance to best industry practices. CRISP-DM model was chosen as DM technique. Business understanding upon Russian E-Commerce companies Wildberries, Citilink, M.Video, Lamoda and Sportmaster (its offline and digital activities, its priorities, needs and goals) was obtained (RQ1 was answered). Data understanding of UGC from Twitter (RQ2 — the best choice of data source for current research) was obtained. Data is collected and prepared for modeling and evaluation stages. During the process of data preparation, it was found out that tweet set for Lamoda on 90+% consists of factual data (coupons sharing and news about company). Only four companies – Wildberries, Citilink, M.Video and Sportmaster – to be analyzed in empirical part. Polarity classification modeling process is schematically explained in Figure 20.

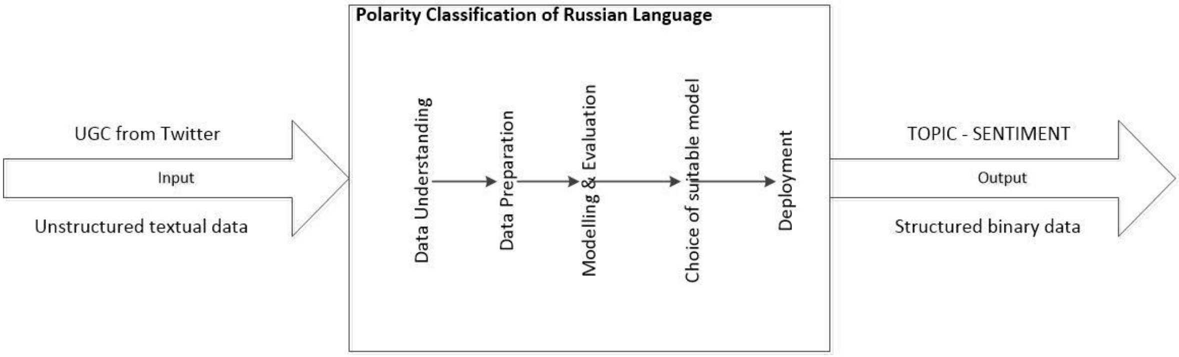


Figure 20. Polarity classification modelling process

Current studies upon dictionary-based and ML/DL-based models of different configurations and their performance baselines were analyzed (RQ3 was answered). Evaluation criteria for each type of models applicable to research purposes were described (RQ4 was answered). The criteria to choose the most relevant model were stated (RQ5 was answered) and final modelling process was shaped.

CHAPTER 3. POLARITY CLASSIFICATION MODELING

Third chapter is concentrated on late stages of modeling, evaluation and deployment of SA model. In previous chapter, training and testing datasets were collected and prepared. In addition, polarity classification models with highest performance baselines were analyzed and the most relevant to research purposes (with regard of such criteria, as researched language and specifics of testing dataset) models were selected. To answer RQ6, these preselected models are built and iteratively improved, until they are superior to existing services for polarity classification of Russian language. Finally, applications (RQ7) for these models are identified and described. Empirical research framework presented in Figure 21.

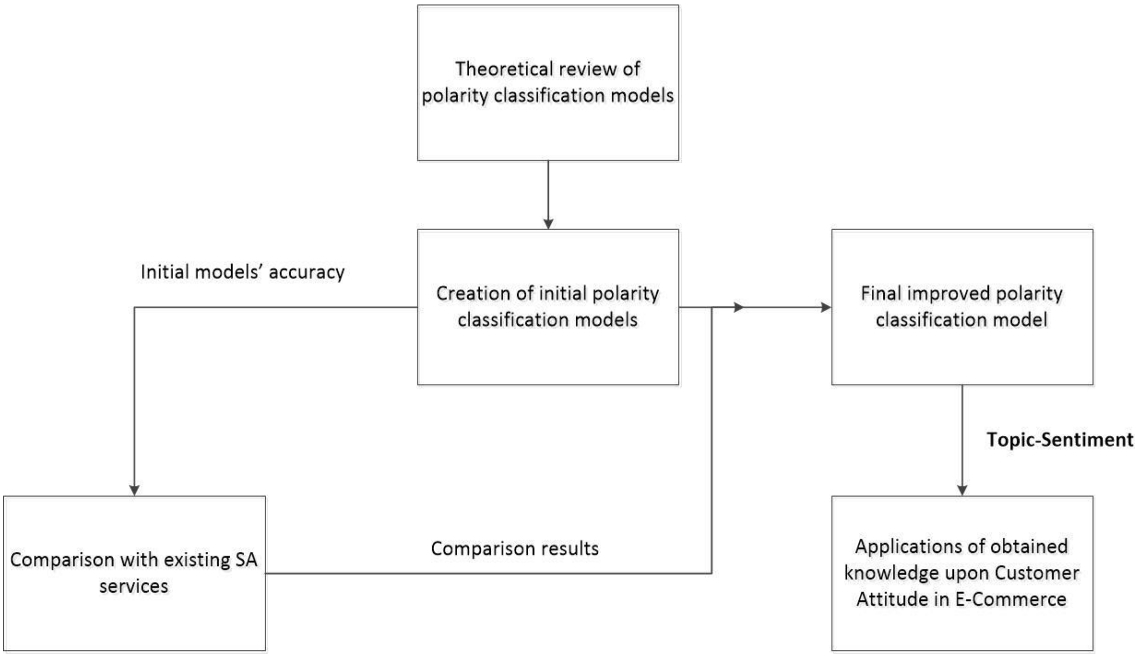

Theoretical review of polarity classification models
Initial models' accuracy
Creation of initial polarity classification models
Final improved polarity classification model
Topic-Sentiment
Comparison with existing SA services
Comparison results
Applications of obtained knowledge upon Customer Attitude in E-Commerce

Figure 21. Empirical research framework

3.1. Initial model creation

Considering type of output desired model should provide manager with (RQ1), model will be enriched with topic modelling functionality. The expected output of the whole model is 'topic-sentiment' pair for every piece in testing tweet corpus. Several based on theoretical review of existing polarity classification research' propositions were made:

1. LDA modeling algorithm is the most suitable for aspect extraction from small datasets with no a priori knowledge upon aspects within dataset (Farhadloo, 2016);
2. Amount and quality of collected training data (19300 entries from Linis-Crowd and RuSentiLex) are sufficient to get F1-score of 50+% for ML-based models on testing tweet dataset (Loukachevitch & Rubtsova, 2016);

3. Amount and quality of collected vector base (30000000 vectors) is sufficient to get F1-score of 50+% for ML-based models on testing tweet dataset (Arkhipenko, 2016);
4. word2vec is the most suitable feature selector for short-text classification tasks (Karpov, 2016; Rubtsova, 2017);
5. Support Vector Machines is the most suitable classifier for polarity classification of Russian language (Loukachevitch & Rubtsova, 2016);
6. F1-score of polarity classification model will exceed F1-score of existing polarity classification services (due to strong impact of Russian language and Twitter specifics).

With these propositions in mind, initial models (both topic modeling and polarity classification) were created and tested.

3.1.1. Topic modelling

Initial simplified LDA model was built with the help of Python script and several Python libraries, such as gensim (gensim, 2018) and scikit-learn (scikit-learn, 2018). Testing datasets for each companies were additionally processed with POS-tagger and only nouns were selected for topic modeling procedure. Features for testing datasets were extracted with TF-IDF tool and those features were analyzed with LDA model. Code for Initial topic modeling may be seen in (Appendix 15. Code for Initial LDA model). Following topics for every company were automatically extracted (see Table 5).

Table 5. Automatically extracted aspects for researched companies in initial model

Company	Aspects
Wildberries	‘Coupons’, ‘Shoes’, ‘Accessories’
Citilink	‘Service’, ‘YouTube’, ‘Guarantees’
M.Video	‘Delivery’, ‘iPhone’, ‘Advertising’
Sportmaster	‘Bonuses’, ‘Bicycle’, ‘Buying’

Those are initial aspects, which are used in initial polarity classification model’ creation. On the next step of initial model’ creation, to each of these topics some polarity value (‘Positive’ or ‘Negative’) is assigned.

3.1.2. Creation of initial polarity classification models

Going along with standard SA modelling process, the first step is to select features. Word2vec and n-gram are used as feature selectors (to test if word2vec increase accuracy in comparison to non-DL feature selectors), due to their high accuracy demonstrated in research by

(Arkhipenko, 2016), (Karpov, 2016) and (Rubtsova, 2017). For word2vec, standard library of vectors with 30mln vectors was initially used. For n-grams, Google N-Grams in Russian (20mln items) were used.

Second step is to assign weights. In case of word2vec, this step was performed automatically on the stage of feature selection. In case of n-gram, values were assigned with the help of SentiWordNet, which has small library of sentiment-labeled Russian words, and partially manually (due to specifics of Twitter paralinguistic content).

Third step is to build classifier. Two most accurate in theory classifiers were tested — Support Vector Machines and Recurrent Neural Networks. Training data for supervised (19300 labeled tweets) and unsupervised (30000 unlabeled texts) learning was used respectively.

Table 6 presents initial performance levels of different variations of supervised ML-based and unsupervised DL-based polarity classification of Russian language' models on testing tweet test (comprised of all researched companies' tweets).

Table 6. Initial polarity classification models' accuracies

Feature + Weight	Model	Precision	Recall	F1-score
n-gram	SVM	0,31	0,38	0,34
word2vec	SVM	0,37	0,51	0,43
word2vec	RNN	0,33	0,47	0,39

Out of all models analyzed, supervised model with word2vec as feature selector and SVM as classifier showed the best accuracy.

3.2. Comparison of initial models with existing services

The next question that pops up is as following: Is it necessary to create and improve the model that fulfills research goal from scratch? Are there any existing services that allow performing polarity classification with the same level of accuracy as analyzed before state-of-the-art models — or at least may be additionally trained to be more domain-specific?

3.2.1. Comparison of initial models' accuracies with implemented services' accuracies

1. Existing theoretical state-of-the-art models show average accuracy (F1-score) higher than 50%. If existing services shows sufficiently higher precision, recall and F1-score for researched companies' UGC, then there is no point in creating new model.

2. Search for existing services for polarity classification of Russian language. As results of queries ('sentiment analysis Russian', 'анализ тональности твитов текста', 'polarity classification Russian', 'определение тона твита', 'анализ мнений потребителей') to Google

and Yandex search engines and analysis of numerous (250+) search results, only 2 services capable to perform polarity classification of Russian language were identified: indico.io and Repustate. Each service declare availability of service for SA of Russian language. To handle their functionality, both services provide access via APIs. Free trials with limited amount of API calls were chosen for research question (RQ6) testing. Limited amount of calls does not constrain functionality and does not affect sentiment classification' accuracy. However, both services has no functionality to be additionally trained or somehow be customized for domain-specific tasks. This means that the obtained result of the initial testing is the best result, which may be obtained at all.

3. In addition to direct search, several specific or not advertised services were analyzed. However, all of them appeared to be inapplicable to polarity classification of Russian language:

- SentiStrength — widely spread SA of English language' tool. However, it is not precise when applied to Russian language due to its morphological specifics — and it has no ability to customize/enrich its power;
- SerpStat — SEO-oriented service, which claims to provide text analysis and sentiment analysis for its users. After registration and overview of service actual capabilities, it turns out that service provides only keyword sentiment analysis, which is not applicable for 1+ word sentences' analysis;
- IBM Watson — comprehensive, highly customizable and powerful tool, which showed high accuracy for English tweets. However, it turned out that it has no internal Russian lexicons and has no way to upload them directly or train the classifier without creating cumbersome pipeline with usage of IBM Cloud, Cloud Floundry Apps and Node-Red, which means Russian language can not be processed and analyzed;
- SentimentMetrics — toolkit for social media monitoring rejected all attempts to register and get access to service functionality and GUI or at least their API;
- iTeco — N-view Object Level Sentiment Analysis program is a part of large and complex software package named Analytical Courier and it is sold only in bundle with tens of other, not always relevant, programs;
- Eureka Engine — accidentally founded proprietary software, which proves to be a part of larger software package. However, it does not seem feasible to learn about existence of this service without direct-targeted search.

4. Accuracy measurement. The same prepared earlier testing tweet corpus was used to benchmark accuracies of different services. The output both services provide was a tweet sentiment evaluated from 0 to 1. Since they work with binary classification, results may be interpreted as negative (0-0.5) and positive (0.51-1). Since testing sample has class imbalance specific, Precision, Recall and F1-score were used as measurement metrics.

3.2.2. Results and conclusions

All existing services were initially tested on English language sample and proved to be functioning correctly (0 for negative, 1 for positive). Then, testing tweet set, combined from all companies' tweets, was analyzed with those services and initial models (Appendix 16. Code for Comparison of initial models with existing services). Results of benchmark are presented in Table 7. In addition, examples of output in (Appendix 17. Polarity classification output by existing services).

Table 7. Existing services vs chosen models accuracy

Service	Precision	Recall	F1-score
Repustate (repustate, 2018)	0	0	**0**
Indico.io (indico.io, 2018)	0,34	1	**0,34**
Eureka Engine (eurekaengine, 2018)	0,33	0,45	**0,38**
Supervised model (word2vec + SVM)	0,37	0,51	**0,43**
Unsupervised model (word2vec + RNN)	0,33	0,47	**0,39**

Several conclusions to be made:

- Not all the services declaring they can perform SA of Russian language can actually perform it — Repustate works well with English language but did not classify Russian language at all;
- Some services are not capable of accurate analysis of Russian language due to its morphological and syntactical specifics (e.g., SentiStrength);
- Some services are not capable of working with unbalanced data, which means they are not be capable of performing real-time monitoring of brand mentions;
- Specifics of UGC (e.g., availability of paralinguistic content) strongly affect general-purpose services' accuracy and requires iterative training on domain-specific corpora;
- The main disadvantage of existing services is in lack of ability to extract aspects. All the services only do sentence-level polarity classification without linking it to specific aspect within the sentence;

- As it was proposed, created during current research initial domain-specific models showed higher F1-score, than existing services.

3.3. Final polarity classification model

3.3.1. Initial model drawbacks

After benchmark it became clear that polarity classification of Russian language' models based upon both supervised and unsupervised ML show better accuracy than existing services, thus, suits better to answer research question (RQ7). After getting initial results, manual verification of both topic extracted and polarity values assigned was performed. The main drawbacks of initial models are:

- Several aspects were chosen incorrectly (e.g., 'YouTube' for Citilink);
- Due to their implicit nature, several significant aspects were not chosen at all;
- Accuracy upon testing tweet dataset was still insufficiently low (in comparison to theoretical models' performance).

3.3.2. Initial models' improvement

Due to implicit nature of the most of topics mentioned in testing tweet dataset, as well due to poor coverage of Russian language by existing topic modeling software. To improve this model programmatically, it is necessary to make changes in algorithm (increase number of learning iterations) and increase amount of data fed to the model. Due to the lack of additional testing data, it was decided not to improve model programmatically, but to extract explicit and implicit topics manually. The topics presented in Table 8 were manually extracted from testing tweet datasets.

Table 8. Manually extracted aspects for researched companies in final model

Company	Aspects
Wildberries	'Advertising', 'Delivery', 'Discounts', 'Exclusivity', 'Location', 'Prices', 'Website'
Citilink	'Advertising', 'Consultants', 'Delivery', 'Return Policy', 'Website'
M.Video	'Delivery', 'Prices', 'Products', 'Discounts', 'Consultants', 'Competitors', 'Website', 'Return Policy', 'Location'
Sportmaster	'Advertising', 'Consultants', 'Competitors', 'Prices', 'Products'

Using word2vec features leads to improvement to results comparing to usage of n-gram (with the same SVM classifier) in more than 3%. It appears that implementation of SVM classifier lead to higher accuracy than implementation of RNN. However, final accuracy of 43% is still not sufficient and is much lower than was initially proposed. It means that some of modeling steps should be iteratively improved.

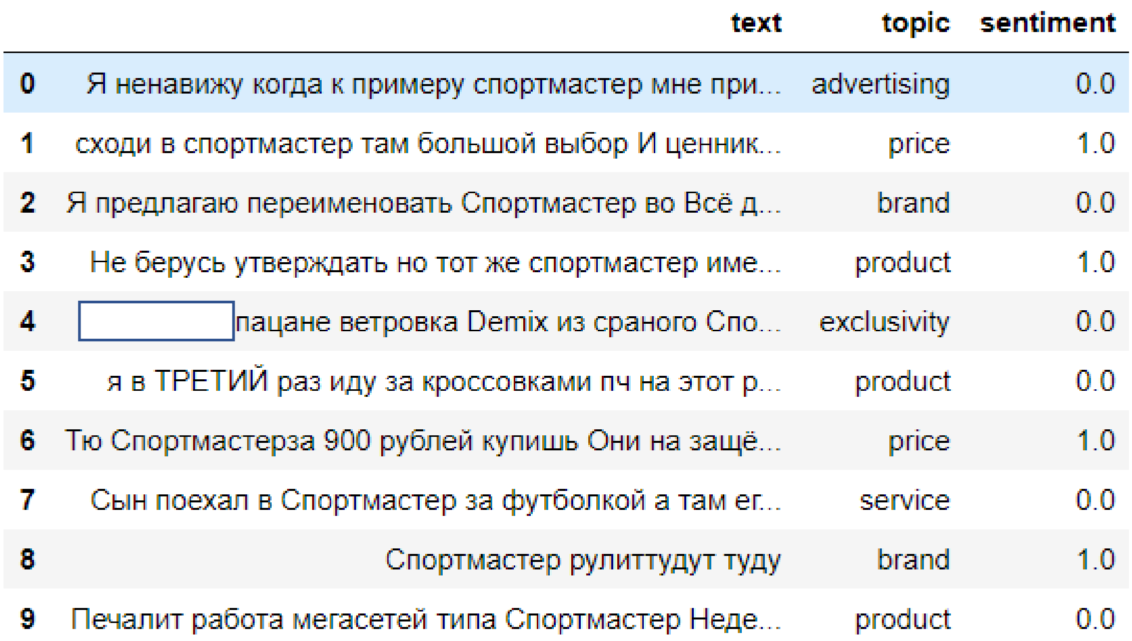
Feature selector was enriched with additional collected words embedding from Russian Distributional Thesaurus (nlpub, 2017). With its size of 12.9bln of words embedding, it should give solid increase in F1-score in comparison to initial model, because efficiency of such approach to vectorization of Russian language was proved by (Arefyev, 2015). Classifier was additionally trained on external library of Russian language tweets, collected by Rubtsova (mokoron, 2017). It consists of 226834 tweets, with balance between negative and positive tweets. The main steps of final model' creation (feature selection and classifier training) may be seen in Python Notebook (described with #). The whole code for final models may be found at author's Github repository (Github, 2018).

3.3.3. Final polarity classification model

Aspect extraction based on LDA algorithm and POS-tagged dataset was tuned in regard to specifics of each company's UGC. The best feature selection type and weight assignment procedure for current research is word2vec enriched with 12.9bln vectors dataset. The best classifier type for current research is Support Vector Machines (outperforms DNN due to specifics of data available and output requirements) trained on 250000 labeled tweets. Iteratively improved DL SA model shows the following accuracy results (F1-score) of 51%. It is represented as Python Notebook, with all the steps (data preparation, feature selection, classifier training, and polarity classification) combined. This final model gave answers on RQ7 and fulfilled the research goal, since it is capable of providing managers in real-time with information about negative or positive attitude towards specific business aspects.

3.4. Application of final model to Russian E-Commerce companies

Final model was deployed and testing tweet data was analyzed. Outcomes consists of labeled testing dataset. Each aspect (topic) within dataset was extracted and labeled with sentiment. Output data was presented in format 'Topic — Sentiment', example may be seen in Figure 22.



	text	topic	sentiment
0	Я ненавижу когда к примеру спортмастер мне при...	advertising	0.0
1	сходи в спортмастер там большой выбор И ценник...	price	1.0
2	Я предлагаю переименовать Спортмастер во Всё д...	brand	0.0
3	Не берусь утверждать но тот же спортмастер име...	product	1.0
4	[redacted] пацане ветровка Demix из сраного Спо...	exclusivity	0.0
5	я в ТРЕТИЙ раз иду за кроссовками пч на этот р...	product	0.0
6	Тю Спортмастерза 900 рублей купишь Они на защё...	price	1.0
7	Сын поехал в Спортмастер за футболкой а там er...	service	0.0
8	Спортмастер рулиттудут туду	brand	1.0
9	Печалит работа мегасетей типа Спортмастер Неде...	product	0.0

Figure 22. Example of output polarity labeled data for Sportmaster

Based on this data, extracted topics were manually analyzed to extract explaining comment upon each aspect. Since positive feedback on SNS has the greatest impact upon sales volumes, positively labeled tweets should be taken into consideration in first place. However, negative feedback provides much larger room for improvement, and gives insights upon problems with different functions.

For Wildberries, out of 86 analyzed tweets seven aspects were identified by model, and most of them with negative sentiment. Results are in Table 9.

Table 9. Extracted Topic — Sentiment for Wildberries

Aspect	Polarity	Comment	Recommendation
Exclusivity	Positive	Unique products that competitors does not have	-
Discounts	Positive	Large number and amounts of discounts, coupons and promo codes	-
Prices	Positive	Low prices in comparison with competitors	-
Advertising	Negative	Excessive and inappropriate e-mail and SMS advertising	Better targeting
Delivery	Negative	Delays and damaged goods	Fix timing issues; change packages
Website	Negative	Poor design of official website	Redesign
Location	Negative	Problems at specific stores	Fix problems

For Citilink, out of 47 analyzed tweets, five aspects were identified by model and all of them with negative polarity. Results are in Table 10.

Table 10. Extracted Topic — Sentiment for Citilink

Aspect	Polarity	Comment	Recommendation
Consultants	Negative	Consultants' absence and their inappropriate behavior	New HR practices
Delivery	Negative	Constant delays and problems with delivered products' conditions	Fix timing issues; change packages
Website	Negative	Bugs in official website	Fix bugs
Advertising	Negative	YouTube advertising caused problems with further video uploading	Stop ads on YouTube until bug is solved
Return policy	Negative	Customer-unfriendly return conditions	Rethink return policy

For M.Video, out of 120 analyzed tweets eight aspects were identified by model, with the most of them with negative sentiment. Results are in Table 11.

Table 11. Extracted Topic — Sentiment for M.Video

Aspect	Polarity	Comment	Recommendation
Products' variety	Positive	Wider choice of products and services than competitors have	-
Discounts	Positive	Large number and amounts of discounts, coupons and promo codes	-
Competitors	Positive	Eldorado return policy	-
Prices	Negative	Strange price changes on old items; prices higher than competitors have	Explain price changes,
Consultants	Negative	Consultants' absence and their inappropriate behavior	New HR practices
Delivery	Negative	Constant delays and courier documentation problems	Fix documentations and timing issues
Website	Negative	Bugs in official website	Fix bugs
Return Policy	Negative	Customer-unfriendly return conditions	Rethink return policy
Location	Negative	Unavailability of several stores; unsatisfying working hours	

For Sportmaster, out of 154 analyzed tweets, six aspects were identified and most of them with negative polarity. Results are in Table 12.

Table 12. Extracted Topic — Sentiment for Sportmaster

Aspect	Polarity	Comment	Recommendation
Prices	Positive	Low prices in comparison with competitors	-
Discounts	Positive	Large number and amounts of discounts, coupons and promo codes	-
Products' variety	Negative	Abundance of mass market products; lack of variety in terms of sizes	Balance product lines
Consultants	Negative	Consultants' annoying behavior	New HR practices
Products' quality	Negative	Low quality products; local brands only	Fix documentations and timing issues
Advertising	Negative	Annoying irrelevant advertising	Better targeting

It can be concluded that in relation to organizational structure and functions within it, obtained labeled data has various possible applications bringing value to researched companies. Various applications are listed in Table 13.

Table 13. SA results' applications

Application/Level	Middle management	Top management
Advertising	1 — Channels' adjustment	2 — Marketing strategy
Customer Service	3 — New scripts and practices	4 — HR policies
Branding	5 — Brand health monitoring	6 — Brand improvement
Pricing	7 — Discounts	8 — Dynamic pricing
Products	9 — Recommendations	10 - Contracts with suppliers
Market research		11 — Competitors analysis
Logistics	12 — New rules and practices	13 — Stores' locations
IT	14 — Content' updates	15 — Website /App improvements
Return Policy	16 — Customer interaction	17 — New policy terms

All these managerial applications are applicable to any large Russian E-Commerce with both online and offline points-of-sales (or self-discharge points), partially being more relevant to companies with tangible products rather than purely service-oriented E-Commerce. Most of these applications (and its value for E-Commerce companies) are described in previous studies upon SA applications for E-Commerce (Qiu et al., 2018; Poecze et al., 2018). Below are elaborations upon final polarity classification model' output applications for E-Commerce companies:

1. Advertising and Social Media Marketing appear to be the most straightforward application of analyzed UGC. With the marketing in general getting more customized and transforming from mass marketing to '1-to-1' paradigm, advertising and SMM are moving in the same direction. For Russian E-Commerce

companies, more targeted and relevant advertising will lead to increase in ad efficiency and decrease in negative perception of companies' marketing efforts.

2. Better understanding of customers allows top-management to alter marketing strategy, for example to tune it in terms of Segmentation (current customers vs. potential customers), Targeting (which segment to focus on in advertising) and Positioning (how customers perceive company's brand and main value propositions).
3. Customer Satisfaction level from interaction with salesperson (e.g., in M.Video and Citilink stores workers are both consultants and salespersons) strongly influence the final purchase decision. Improved scripts and personal trainings may improve efficiency of store employees.
4. HR policies may be used as a strategic tool of changing the format of interaction between store workers and customers on all levels and not in specific points-of-sales, but on the firm level.
5. Brand health monitoring allows company to avoid reputational damages and act proactively in case of consumers' rage against the company. Due to the direct influence of brand comprehension on customers' purchase intentions, brand health monitoring is an important procedure for every E-Commerce manager.
6. Brand policies, which may start at simple change of advertising messages to complete rebranding, may be guided by customer desires, extracted by polarity classification model. Simple changes in messages' content along with its real-time Customers' Attitudes analytics may lead to increase in brand loyalty.
7. Sometimes discounts are more efficient and appropriate, and understanding of Customers' Attitudes help to identify this timing correctly. Along with increases in sales volumes, additional loyal customers may be obtained.
8. Dynamic pricing is the most applied tool for increasing revenues without losing customers or going for sales. In addition, since price is very important for customer in moment of overt search and alternatives evaluation, it is the best way to outperform competitors.
9. More in-depth knowledge upon customers' desires to buy specific products or specific brands along with will lead to more relevant recommendations. For example, M.Video in 2017 launched a pilot project of Amazon-like upselling recommendation algorithm with e-mail sender (ComNews, 2017). Created polarity classification model may help to enhance current system.

10. Contracts with suppliers may lead to change of product lines to ones more desired by customers. Higher the demand for products on shelf, higher the sales. Extra quality control is also needed to assure high level of customers' satisfaction, which is direct influencer for purchase intention.
11. Competitors analysis may help to exploit some marketing activities or operational-related approaches for the sake of researched companies' striving.
12. Fails in delivery worsen the whole experience of customer-company interaction. Fixing these problems is a feasible task, especially if delivery service is in-house.
13. Experiments with stores' locations (numerous customers complained about their inconvenient location) or implementation of more advanced delivery system (e.g., Amazon's Prime) are applicable to fix current issue.
14. Often, because of out-of-date information on website, customers get negative experience. In addition, the absence of relevant information lead to decline in sales and negative brand perception.
15. Websites and apps have a tendency to become out-of-date in comparison to highly dynamic online competitors, which negatively affects customers' desire to buy online via this shop.
16. If customers are not happy with initial purchase, store managers should fix this issue so at the end customer is satisfied with his purchase and is eager to repeat it in future.
17. In case of numerous complaints upon return policy, some part of it may be reshaped or improved to make customers more eager to buy and not be afraid of fraud from company's side.

3.5. Summary of chapter 3

The main goal of chapter 3 was to create polarity classification model, which provides valuable for E-Commerce companies' managers output. In previous chapter, different theoretical polarity classification models were analyzed in terms of their performance for Russian language short-text data. In chapter 3, those initial models were created and trained on Big Data text corpora. These initial models were upgraded with ability of topic modelling, since only this configuration may bring to decision maker useful and actionable information. Then, initial models' accuracy was compared with existing commercial solutions for SA of Russian language services on the same testing dataset (Russian E-Commerce companies' related tweets), with created in current research models showing higher accuracy than existing services. Accuracy results measured and the most relevant polarity classification model was picked with F1-score as main measurement metric (final PR AUC metrics). After that, model results were manually verified and main drawbacks were identified. The most accurate initial model was improved — topic modeling algorithm was changed and polarity classification model' elements were additionally trained (feature selector + classifier). The final model setup consists of word2vec feature selector and SVM classifier. The final accuracy of model is 51% (again, F1-score) and it is higher than of all initial models. The output in form of 'Topic — Sentiment' was obtained after application of created final model to testing data with tweets containing indirect mentions of four Russian E-Commerce companies (Wildberries, Citilink, M.Video, Sportmaster). Testing tweet set was analyzed and Customers' Attitudes towards different aspects of companies' activities were extracted and labeled. Domain-specific recommendations for Russian E-Commerce companies upon advertising, marketing mix, product-related features and customer service were given.

Discussion and conclusions

Discussion of the findings

Current study was devoted to application of Sentiment Analysis tasks to assess customers' attitude towards researched company. Research gap was in the lack of empirical studies upon applications of polarity classification of Russian language that are beneficial for managers of E-Commerce companies. The main research goal of this study was in creation and test of polarity classification model, which allows managers of Russian E-Commerce companies to achieve better understanding of customers' attitudes towards their brand, products and services. The goal was achieved through theoretical review of existing literature on the topic of Sentiment Analysis of Big Data with in-depth overview of studies upon different polarity classification models and methods of obtaining knowledge upon customers' attitudes with their assistance. On the basis of theoretical review, polarity classification model was built, compared with existing services and tailored to specifics of current research setup. To achieve research goal, seven research questions were stated.

The first research question (RQ1) was to understand type and format of sentiment labeled data, which may be useful for managers in E-Commerce. Since simple polarity identification alone is not enough to get sufficient for output some valuable for business data, additional knowledge in terms of automatically extracted topics was gathered. Final output is represented in format 'Topic — Sentiment', for example, 'Price — Negative'. This knowledge allows managers to promptly identify possible threats to company and handle those issues. Five companies were chosen as cases of Russian E-Commerce companies interested in obtaining such type of output — Wildberries, Citilink, M.Video, Lamoda and Sportmaster. When switching specifically to E-Commerce companies, numerous applications in advertising, customer service and marketing mix management were identified. With advanced knowledge of customer, managers of E-Commerce companies may enhance marketing activities through tuning advertising channels and messages' content; improve customer service with more quick and valuable for user complaint handling; improve marketing mix via rethinking current targeting and positioning directions. For more detailed information upon possible applications of this type of sentiment labeled data, see ongoing 'Managerial implications' section.

The second research question (RQ2) was to find relevant UGC sources and how to extract data upon Customer Attitude (which is needed to build polarity classification model) from it. Source criteria, such as company's presence at SNS (community, page, group, etc.), ability to access and lack of moderation from company side, were stated. Since SNS are currently the main source of UGC, different popular within Russian audience SNS were reviewed. Among such networks, as VKontakte, Odnoklassniki, Facebook and Twitter, the latter one as the most appropriate source of UGC was selected due to its suitability to stated source criteria.

The third research question (RQ3) was to analyze performance baselines for state-of-the-art polarity classification for Russian, English and multilingual models. Studies in both English and Russian languages, as well as multilingual studies were reviewed. Both lexicon-based and Machine Learning based approaches towards polarity classification were overviewed. The most efficient models appears to be ML and DL-based models with n-grams and word2vec as feature selectors, and SVM and RNN as classifiers. The most efficient models from oriented on Russian language studies were later created and tested in empirical part.

The forth research questions (RQ4) was in selection of appropriate accuracy metrics. Those metrics should be aimed on accuracy measurement of classifier performance, which is tested upon imbalanced data. The most commonly used in industry ROC AUC and PR AUC metrics were compared, with PR AUC (F1-score, specifically) selected due to specifics of testing data used in research.

The fifth research question (RQ5) was to identify the most relevant polarity classification of Russian language' model basis for research goal and specific industry (Russian E-Commerce). Concerning desired output, final polarity classification model should be able to extract aspects upon researched companies, to handle specific for Twitter messages paralinguistic content and misspellings, to classify polarity correctly with accuracy heading to state-of-the-art performance baselines. In current research, out of all reviewed models, the most relevant polarity classification model' setup is word2vec as feature selector and SVM as classifier.

The sixth research question (RQ6) was to analyze performance baselines of existing general-use proprietary services of polarity classification of Russian language. At first, numerous services were found and questioned on their ability to handle Russian language' analysis. Then, three suitable services — indico.io, Repustate and Eureka Engine were selected. To perform this analysis, the same testing dataset with tweets upon selected Russian E-Commerce companies' was used. Existing services (functionality was accessed via APIs), and created during research initial models were tested with this dataset. Since imbalance of testing dataset, PR AUC metric (F1-score) was used to assess accuracy. Roughly speaking, F1-score calculates ratio of correctly and incorrectly predicted sentiment values. It turned out, that existing services does not show accuracy higher than 38%, while theoretical models for Russian language showed 50+% accuracy and initial models in current research demonstrated 40% accuracy. Along with lack of ability to customize existing services, it was decided to create tailored for researched companies' model through improving initial models.

The last research question (RQ7) related to the process of creation tailored for researched companies' polarity classification of Russian language' model aimed on assessment of Customer Attitude in UGC from SNS. First, several propositions based on theoretical review were made — upon efficiency of separate elements of polarity classification models (feature selector and classifier) and accuracy of initial models vs existing SA services. Initial models for topic modeling and polarity classification were created and tested. Both extracted topics and labeled sentiments were manually verified and were not sufficiently accurate to fulfill research goal. Several iterations and improvements in topic modeling algorithm as well as in polarity classification algorithm were made — amount of words embedding for feature classifier was increased to 12.9 bln and volume of training data for classifier was increased to 250000 tweets. The final model was created and tested upon testing tweet dataset, which consists of indirect mentions of researched companies (Wildberries, Citilink, M.Video and Sportmaster) during 2018 and may be seen in Figure 23.

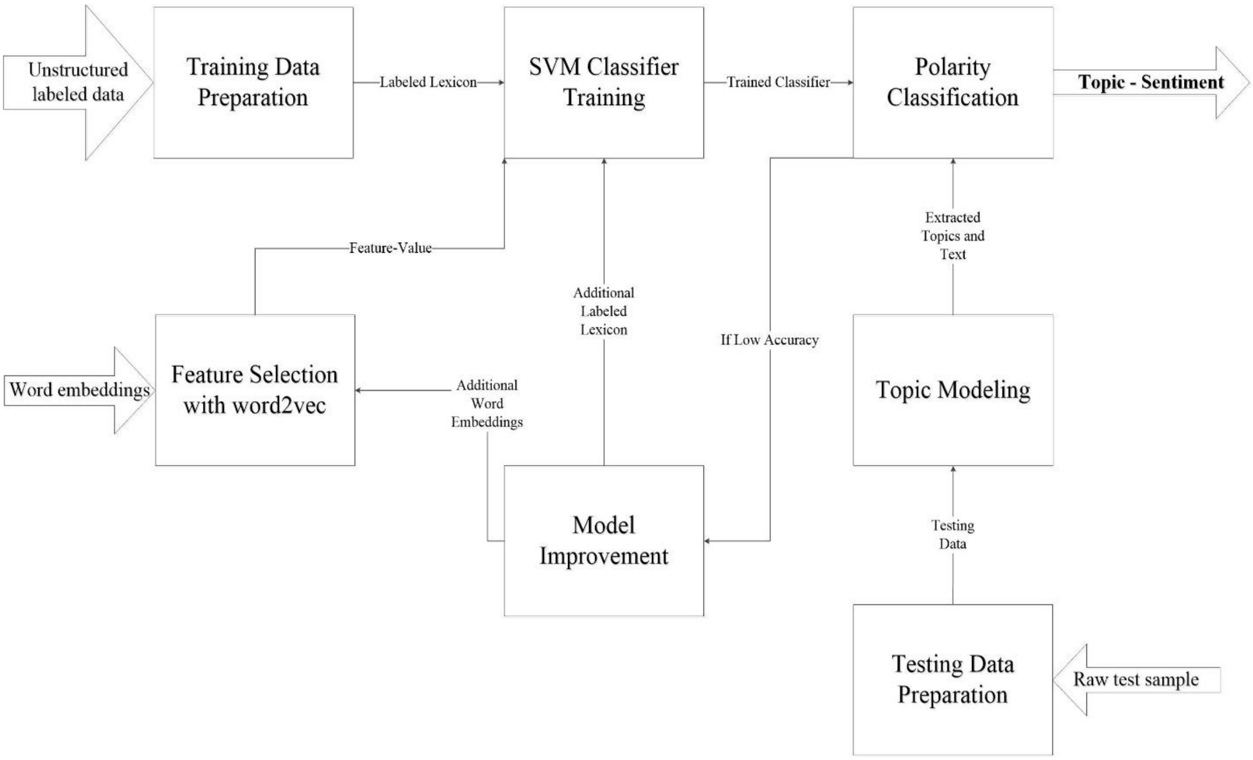


Figure 23. Final polarity classification model' framework

The final model consists of five main steps: data preparation (data collection and data processing), feature selection, classifier training, topic modeling and polarity classification. In this thesis, all the steps are tailored to research specifics, ready to be used by manager (code for final polarity classification model in Python Notebook is posted at author's Github repository (Github, 2018)) and are described in details below.

Step of data preparation include collection and processing of training and testing data. For data collection, Twitter application' credentials are obtained at first. Credentials were obtained via straightforward login at Twitter Apps (Twitter, 2018). Then, there are two different approaches to collect data — in real-time manner or in historical manner. In case of real-time, database to store all necessary data is created and data collection process is pipelined. Database is designed on the base of NoSQL CloudantDB, and pipeline flow is made with Node-RED. To perform real-time data collection, manager have to create Twitter app, create database (SQL or NoSQL — depends on managers' preferences) and pipeline data collection flow. To do this, manager should have understanding of how databases work and be familiar with basics of JavaScript. In case of historical data collection, Python script to gather data in necessary for further modeling format (.csv or .json) is written. To perform historical data collection, manager have to create Twitter app and launch script in with obtained Twitter credentials ('consumer' and 'app' keys). To do this, manager should be familiar with basics of Python language. Output of this step — raw collected training and testing datasets.

Data processing is performed after obtaining understanding of needed data output. It include training and testing data processing, as well as separation (or merge) of datasets into needed ones. It is performed within Python Notebook with the help of such libraries, as python-twitter, pandas and NLTK. To perform data preparation, manager have to put input data file into Notebook and launch the code. To do this, manager should be familiar with basics of Python language. Since quality of data directly affects final model accuracy, in this thesis was made an attempt to find the most suitable for short Russian texts. In comparison to studies by (Rubtsova, 2017), data was prepared specifically for business-oriented problems — separate POS-tagged dataset for topic modeling for each company, morphology. Output of this step — prepared and structured training and testing datasets.

Step of feature selection include gathering of feature vectors and training word2vec model on them. Feature vectors may be gathered manually from domain-specific tweet corpora, or collected from external open-source text corpora (such as (Rusvectores, 2018)). To gather and prepare domain-specific corpora, manager have to use the same tools, as for abovementioned data preparation' tasks (with the same required skillset). To collect external corpora, manager have to visit the website and download corpora (no special skills required). Feature selector' training is performed within Python Notebook with the help of such libraries, as pandas, gensim and word2vec. To train feature vectors, manager have to input data file into Notebook and launch the code. To do this, manager should be familiar with basics of Python language. Output of this step — prepared for classifier training features. The main contribution of current thesis to step of feature selection is in selection of the most suitable for short texts feature selector, which is word2vec, and training vector dataset sufficient to achieve accuracy of 50+%, which is combined dataset of RusVectores and tweets' vectors.

Step of classifier training is aimed on building a model, capable to separate positive features from negative as accurate as possible. It is performed within Python Notebook with the help of such libraries, as pandas and scikit-learn. To train classifier, manager have to input data file into Notebook and launch the code. To do this, manager should be familiar with basics of Python language. Output of this step — prepared for classifier training features. The main contribution of current thesis to step of classifier training is in selection of the most suitable and accurate for Russian language' classifier, which is Support Vector Machines.

Step of topic modeling include extraction of topics and accuracy evaluation. It is performed within Python Notebook with the help of such libraries, as pandas, scikit-learn and gensim. To extract topics from necessary dataset, manager have to input prepared for topic modeling data file (see 'data processing' section) into Notebook and launch the code. To do this, manager should be familiar with basics of Python language. Output of this step — extracted and ready-to-be-labeled topics for each company.

Step of polarity classification include classification of testing data and accuracy evaluation. It is performed within Python Notebook with the help of such libraries, as pandas and scikit-learn. To mark with sentiment extracted topics, manager have to input ready-to-be-labeled topics along with testing dataset into Notebook and launch the code. To do this, manager should be familiar with basics of Python language. Output of this step — list of sentiment-labeled topics for each company. The main contribution of current thesis to current step is in creation of combined Russian language' polarity classification model, capable to outperform existing SA services.

The entire model in current thesis consists of two parts. The first one is created for data collection and is represented as pipeline of Node-Red with Cloudant database (data is grabbed from Twitter with Node-Red and stored at database). The second one is for all data preparation and modeling procedures and is represented as a Python Notebook (.ipynb format). To use it for her purposes, E-Commerce company manager should be familiar to basics of Python and JavaScript languages programming (with additional knowledge of such libraries, as pandas and scikit-learn), able to read and analyze other developers' code, be familiar with how SQL databases work and capable to launch program via convenient to him command line or IDE. For managers' convenience, the whole model may be translated into stand-alone web service (with some help from professional Python developers), with access via API and graphical user interface.

After the manual verification of sentiments' labeled to testing data, it may be concluded that final model was tested successfully — the most of explicit polarity expressions were labeled correctly, with main mistakes made in sarcasm detection, slang misinterpretation and numerous negations handling.

Thus, it can be concluded that all research questions of the thesis were answered, thus, fulfilling stated research goal.

Theoretical contributions

Current research initializes research upon applications of polarity classification of Russian language in business, specifically in E-Commerce as well as elaborate upon topics of collection of Big Data from Social Networks and extraction of customers' attitudes out of this unstructured textual data. Talking about Big Data, its variety is referred primarily. Data from Social Networks includes different types of textual data (text, symbols, emoticons, hashtags, etc.) and is written in various languages (e.g., for Lamoda, several reviews in English were found). Along with variety, volume is important. Amount of textual data in Social Networks is increasing daily (Statista, 2018), and, though in current thesis quite small amount of data was used (250000 tweets as training data and 400 tweets as testing data), research of Customers' Attitudes from User-Generated Content is getting more and more precise and insightful due to emergence of these large datasets.

The main theoretical contribution was made towards theory upon creation of polarity classification models. Advantages and disadvantages of different polarity classification of Russian language models' were analyzed. Various polarity classifications models with different setups were analyzed and described in terms of their applicability for research goal. It was proved that the most relevant for such tasks feature selector is word2vec and classifier is SVM, in line with results obtained by (Loukachevitch & Rubtsova, 2016). Created model may serve as a baseline for further business-related polarity classification of Russian language' models.

Since the lack of literature upon applications of polarity classification models in E-Commerce, current research provides a basis for more in-depth research upon polarity classification of Russian language in business. It provides theoretical review of the polarity classification modeling process specifically for Russian language short-text messages with focus on practical managerial implications. It describes the process of data mining concerning business-oriented tasks and goes in details in process of data preparation of Twitter data in Russian language. It does in-depth analysis of existing approaches to polarity classification and compare different setups for polarity classification models. In further research, it may be used as a platform for more technically sophisticated (with larger training datasets and more advanced classifier algorithms) Customers' Attitudes analysis.

Managerial implications

Along with theoretical contribution, various managerial implications were provided. One important notice here: since the final built polarity classification model has one straightforward managerial implication — extract knowledge upon Customers' Attitudes from User-Generated Content, the focus of 'Managerial Implications' part is on the aspects of how this extracted knowledge may be applied in Russian E-Commerce companies.

The main managerial implication is quite straightforward – the research goal was fulfilled and functioning polarity classification model was created. It means that E-Commerce companies' managers can use the final model to get 'Topic-Sentiment' for their own data, which may include UGC in Russian language collected from different sources (however, the best results will be made on short-text Twitter-alike type of data).

Since companies are interested in getting more proactive in company-customer interaction, knowledge upon Customers' Attitudes is becoming crucial for companies' success. During the analysis of User-Generated Content made by customers of the largest Russian E-Commerce companies, numerous important for customers aspects were identified. There are several functions of company, extensively commented by customers — advertising, customer service, branding, pricing, products' variety and quality, logistics, IT and others. Managerial implications within all these functions may be divided into groups based on level of management who will use the output of polarity classification model. For top management, possible applications include adjustment of marketing strategy, improvement of HR policies, brand image improvement, implementation of dynamic pricing, restructuration of contracts with suppliers and competitors' analysis. For middle management, possible applications include adjustment of advertising channels and content, implementation of brand health monitoring, new standards of interaction with customer online and offline and more efficient handling of returns and complaints.

Limitations of the study

First, numerous model' improvements lay beyond scope of current Master Thesis and it was decided to resort to more straight and simple techniques to create the model. Since this study is business-oriented in the first place, more attention was paid to managerial implications and benefits for business rather than to scrupulous descriptions of technical details.

Second, limited amount of data upon researched company. Due to the specifics of Twitter APIs, the amount of data available for research is limited to data gathered through real-time Streaming API. Another reason for tweet number shortage is limited popularity of Twitter among Russian speakers. In case of large Western companies with socially English-speaking active customer base, there is no problem of sufficient data collection arising. For example, there are about 5000 tweets per day directly or indirectly mentioning Starbucks, while there are only 5 tweets per day directly or indirectly mentioning M.Video. A good solution may be to find other sources of UGC, which currently are not moderated by companies (e.g., Twitter is moderated by company if it has functioning account or page, interacts with company's followers and handles direct mentions). For example, useful information on companies may be found in review sections at websites of web aggregators (such as Yandex Market). Besides, there is lack of publicly available datasets, which may be used as training data.

Third, along with its limited amount, data is also scarce in terms of amount and diversity of data sources. Due to different upcoming changes in terms and conditions of various SNS and open forums towards usage of data by third-parties (caused, for example, by abovementioned Cambridge Analytica data scandal) and restrictive actions of Russian legislative and pro-government organizations (e.g., Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications), data extraction became more cumbersome procedure.

Forth, since this research in business-oriented, it was decided to 'cut corners' on the topics of linguistics and consumer psychology. While both those themes are directly related to researched topic (extraction of customer thoughts from textual pieces), they were touched upon briefly in chapter 1 (basic elements influencing purchase decision process) and chapter 2 (knowledge of Russian language' morphology needed to preprocess data correctly).

Prospects for future research

First, usage of other types of Twitter data for managerial implications. Along with textual data, Twitter also contains other types of data. Geospatial data, which may be used to perform precise segmentation and solve problems on the level of individual stores. Numerical data upon user behavior on SNS (number of followers, number of followings, number of tweets), which may be used to identify users relationships, influencers and patterns of information spread within these internal 'friendship' networks. Attached images and videos, which usually either show the Customers' Attitude or show some real-life situation within stores, may be analyzed to gain additional information upon customers. Meta-level features can be extracted for the same purposes.

Second, current model can be improved with more technically complex and sophisticated approaches to SA modeling steps. For lexicon-based polarity classification, overall accuracy of model may be improved with integration to training data more wide or domain-specific dictionaries. In addition, lexicons with high results in numerous studies, such as MPQA, Bing Liu’s and LIWC, may be translated and used as training data. For ML/DL-based polarity classification models (since they proved to be the most accurate ones), every step of modelling process may be improved. On the step of feature selection, larger training data corpus (e.g., closed for private use such as (Russian National Corpus, 2018)) may be used to get the higher accuracy. For ensemble models, additional rules may be added to make feature selection more precise. On the step of weight assignment, sentiment labeled lexicons, such as Twitter Sentiment Analysis Dataset, may be translated and integrated into training corpora. Preemptive subjectivity classification may be improved and learned to not only separate tweets into subjective and objective, but also to mine opinions from objective tweets and to test whether they may be useful for managers. Aspect extraction may be improved and identify not only aspects themselves, but also summarize (or cluster and summarize) the Customers’ Attitudes upon selected aspects automatically.

Third, with deeper research upon consumer behavior online and customers’ psychographic characteristics, it is possible to advance SA and retrieve more relevant and sophisticated data upon customers’ shopping intentions and motives. With more profound psychology-based approach, consumer behavior may be interpreted in a more relevant manner. For example, with better understanding of interrelations between building trust and increase of purchase intention, it is possible to use SA as a tool to measure customers’ trust towards brand and find ways to enhance it.

Fourth, there are number of interesting for research topics within Sentiment Analysis, which are strongly related to polarity classification itself. For example, fake opinion’ detection, slang preprocessing and automatic handling of grammatical errors (Tubishat et al., 2018). The most problematic are related to extraction of implicit data. Theoretically, all this implicit data may be extracted with more advanced approach to topic modeling. Along with this, such entities within text as sarcasm and hate may be detected more precisely and be interpreted in a more correct way.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). Sentiment Analysis of Twitter Data. In: Proceedings of the Workshop on Languages in Social Media.
2. Aguwa, C. (2017). Modeling of fuzzy-based voice of customer for business decision analytics. Knowledge-Based Systems, vol. 125: 136-145.
3. AITC. (2017). M.Video online sales showed record growth in the last two years. Retrieved from http://www.akit.ru/онлайн-продажи-м-видео-показали-рек/.
4. AITC. (2018). Russian E-Commerce market volume exceeded 1 trillion rubles. Retrieved from http://www.akit.ru/оборот-российского-рынка-интернет-ри/
5. Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research Trends on Big Data in Marketing: A text mining and topic modeling based literature. European Research on Management and Business Economics, vol. 24(1-7).
6. Anastasyev, D., Andrianov, A., Indenbom, E. (2017). Part-of-Speech tagging with rich language description. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017”.
7. Arefyev, N. 2015. Evaluating Three Corpus-based Semantic Similarity Systems for Russian. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2015”.
8. Arkhipenko, K., Kozlov, I., Trofimovich, J., Skorniakov, K., Gomzin, A., Turdakov, D. (2016). Comparison of neural network architectures for sentiment analysis of Russian tweets. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”.
9. Baccianella, S., Esuli, A., Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC ’10.
10. Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of online consumer reviews: Readers' objectives and review cues. International Journal of Electronic Commerce, 17(2): 99-126.
11. Bahtar, A., Muda, M. (2016). The Impact of User – Generated Content (UGC) on Product Reviews towards Online Purchasing – A Conceptual Framework. Procedia Economics and Finance, vol. 37: 337-342.
12. Balahur, A., & Perea-Ortega, J. M. (2015). Sentiment analysis system adaptation for multilingual processing: The case of tweets. Information Processing & Management, 51(4): 547-556.

13. Bao, H., Li, Q., Liao, S. S., Song, S., & Gao, H. (2013). A new temporal and social PMF-based method to predict users' interests in micro-blogging. Decision Support Systems, 55(3): 698-709.

14. Barbosa, L., Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING ’10: 36-44.

15. Benko, V., Zakharov, V. (2016). Very Large Russian Corpora: New Opportunities and New Challenges. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”.

16. Bifet, A., Holmes, G., Pfahringer, B., & Gavalda, R. (2011). Detecting sentiment change in Twitter streaming data.

17. Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., Haruechaiyasak, C. (2012). Discovering consumer insight from twitter via sentiment analysis. Journal of Universal Computer Science, 18(8): 973-992.

18. Chang, V. (2017). A proposed social network analysis platform for big data analytics. Technological Forecasting and Social Change.

19. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (1999). CRISP-DM 1.0. 1-76.

20. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275: 314-347.

21. Chetviorkin, I., Loukachevitch, N. (2013). Evaluating Sentiment Analysis Systems in Russian. In: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing: 12-17.

22. ComNews. (2017). Retrieved from https://www.comnews.ru/digital-economy/content/110448/opinions/2017-11-13/mvideo-onlayn-blokcheyn-i-made-russia.

23. Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications, 36(3): 6127-6134.

24. Data Insight. (2017). Top-100 Russian E-Commerce companies. Retrieved from http://datainsight.ru/top100/.

25. Dean, J. (2014). Big Data, Data Mining and Machine Learning. New Jersey, Wiley & Sons.

26. Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. Decision Support Systems, 94: 65-76.

27. Devika, M., Sunitha, C., Ganesh, A. (2016). Sentiment Analysis: A comparative study on different approaches. Procedia Computer Science, vol. 87: 44-49.

28. Dialog-21. (2012-2017). Retrieved from http://www.dialog-21.ru/.

29. Ding, A., Li, S., Chatterjee, P. (2015). Learning User Real-Time Intent for Optimal Dynamic Web Page Transformation. Information Systems Research, vol. 26, issue 2: 339-359.

30. Du, J., Xu, H., & Huang, X. (2014). Box office prediction based on microblog. Expert Systems with Applications, 41(4): 1680-1689.

31. Dubatovka, A., Kurochkin, Yu., Mikhailova, E. (2016). Automatic Generation of the Domain-Specific Sentiment Russian Dictionaries. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016".

32. E-Commerce Foundation. (2016). Russia B2C E-Commerce Report. Retrieved from http://www.ecommercefoundation.org/.

33. Engel, J. F., Blackwell, R. D., & Miniard, P. W. (1995). Consumer behavior, 8th. New York: Dryder.

34. Ermakov, A. (2009). Knowledge extraction from text and its processing: current state and prospects. Information technologies, (7):50–55.

35. Eurekaengine (2018). Retrieved from http://eurekaengine.ru/ru/demo/.

36. Fang X., Hu, P., Li, Z., Tsai, W. (2013). Predicting Adoption Probabilities in Social Networks. Information Systems Research 24(1): 1-56.

37. Farhadloo, M., Patterson, R., Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. Decision Support Systems, vol. 90: 1-11.

38. Fersini, E., Messina, E., & Pozzi, F. A. (2016). Expressive signals in social media languages to improve polarity detection. Information Processing & Management, 52(1): 20-35.

39. FSSS. (2017). Russia in figures, 2017. Retrieved from http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/en/main/.

40. García-Cumbreras, M. Á., Montejo-Ráez, A., & Díaz-Galiano, M. C. (2013). Pessimists and optimists: Improving collaborative filtering through sentiment analysis. Expert Systems with Applications, 40(17): 6758-6765.

41. Gensim. (2018). Retrieved from https://radimrehurek.com/gensim/index.html/.

42. Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. Decision support systems, 57: 212-223.

43. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A. (2015). SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15: 470-478.

44. Github. (2018). Polarity Classification Model (word2vec + SVM) source code. Retrieved from https://github.com/alextimakov/Polarity_Classification_Russian.

45. Hagen, C., Ciobo, M., Wall, D., Yadav, A., Khan, K., Miller, J., Evans, H. (2013). Big data and the Creative Destruction of Today's Business Models. Kearney Publishing, Chicago: 1-18.

46. Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. Journal of the Academy of Marketing Science, 43(3): 375-394.

47. Homburg, C., Ehm, L., Artz, M. (2015). Measuring and Managing Consumer Sentiment in an Online Community Environment. Journal of Marketing Research, 52(5): 629-641.

48. Howard, J. A., & Sheth, J. N. (1969). The theory of buyer behavior (No. 658.834 H6).

49. IBM Cloud: Bluemix. (2017). IBM.com. Retrieved from https://www.ibm.com/cloud-computing/bluemix/what-is-bluemix.

50. Indico.io. (2018). Retrieved from https://indico.io/product/.

51. J. Bollen, H. Mao, X. Zeng. (2012). Twitter mood predicts the stock market. Journal of Computer Science, vol. 2: 1-8.

52. Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.

53. Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications, 41(4): 1041-1050.

54. Kao, A., Poteet, S. (2006). Natural Language Processing and Text Mining. USA, Springer.

55. Karavdic, M., Gregory, G. (2005). Integrating e-commerce into existing export marketing theories: A contingency model. Marketing Theory, vol. 5, issue 1: 75 – 104.

56. Karpov, I., Kozhevnikov, M., Kazorin, V., Nemov, N. (2016). Entity based sentiment analysis using syntax patterns and convolutional neural networks. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016".

57. Kazennikov, A. (2017). Part-of-Speech Tagging: The Power of the Linear SVM-based Filtration Method for Russian Language. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017".

58. Keshavarz, H., Abadeh, M. (2017). Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. Knowledge-Based Systems, vol. 122: 1-16.

59. Khabibrakhimov, E. (2018). M.Video acquired Eldorado for 45.5bln rubles. Retrieved from https://vc.ru/37081-m-video-zakryla-sdelku-po-pokupke-eldorado-za-45-5-mlrd-rubley

60. Kirilenko, A., Stepchenkova, S. (2017). Sochi 2014 Olympics on Twitter: Perspectives of hosts and guests. Tourism Management, vol. 63: 54-65.

61. Klekovkina M., Kotelnikov E. (2012). The automatic sentiment text classification method based on emotional vocabulary. Russian Digital Libraries Journal: 118–123.

62. Kotelnikov, E. (2012). Combined method of text tonality automatic identifying. Journal of Software products and systems, no. 3: 189-195.

63. Kotelnikov, E. V., Bushmeleva, N. A., Razova, E. V., Peskisheva, T. A., & Pletneva, M. V. (2016). Manually created sentiment lexicons: research and development. Computational Linguistics and Intellectual Technologies, 15(22): 281-295.

64. Kouloumpis, E., Wilson, T., Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! In: Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM '11: 538-541.

65. Kravchenko, A. (2012). Mining for Opinions Across Domains: A Cross-Language Study. In: Proceedings of the First International Workshop on Sentiment Discovery from Affective Data (SDAD 2012): 67-74.

66. Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P. (2016). From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. Journal of Marketing, 80(1): 7-25.

67. Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70).

68. Lantos, G. P. (2010). Consumer behavior in action. ME Sharpe.

69. Li, X., Wu, C., & Mai, F. (2018). The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis. Information & Management.

70. Li, Y. M., & Li, T. Y. (2013). Deriving market intelligence from microblogs. Decision Support Systems, 55(1): 206-217.

71. Li, Y. M., & Shiu, Y. L. (2012). A diffusion mechanism for social advertising over microblogs. Decision Support Systems, 54(1): 9-22.

72. Liu, B. (2010). Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca.

73. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool, San Rafael, CA.

74. Loukachevitch, N., Blinov, P., Kotelnikov, E., Rubtsova, Y., Ivanov, V., Tutubalina, E. (2015). SentiRuEval: testing object-oriented sentiment analysis systems in Russian. In: Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015, vol. 2: 2-13.

75. Loukachevitch, N., Rubtsova, Y. (2016). SentiRuEval-2016. Overcoming time gap and data sparsity in tweet sentiment analysis. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”.

76. Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. Journal of Marketing, 77(1): 87-103.

77. Mantyla, M., Graziotin, D., Kuutila, M. (2018). The evolution of sentiment analysis – a review of researched topics, venues, and top cited papers. Computer Science Review, vol. 27: 16-32.

78. Mayer-Schönberger, V., & Cukier, K. (2014). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston, Mariner Books.

79. Mazurova, M. (2016). Grammatical Dictionary Generation Using Machine Learning Methods. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016".

80. McKinsey & Company. (2016). The age of analytics: Competing in a data-driven world. McKinsey Global Institute.

81. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4): 1093-1113.

82. Michael, K., Miller, K. (2013). Big Data: New Opportunities and New Challenges. Computer, vol. 46, issue 6: 22-24.

83. Mickiewicz, K. (2016). Manager in digital economy: the evolution of managerial role in organizational success. Journal of Management and Organization, 18(4): 56-68.

84. Mohammad, S.M., Kiritchenko, S., Sobhani, P, Zhu, X., Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16: 31-41.

85. Mohapatra, S. (2012). E-Commerce Strategy: Text and Cases. Springer Science and Business Media.

86. Mokoron. (2017). Retrieved from http://study.mokoron.com/.

87. MongoDB White Paper. (2016). Big Data: Examples and Guidelines for the Enterprise Decision Maker.

88. Morabito, V. (2015). Big data and analytics: strategic and organizational impacts. Springer International Publishing, Switzerland.

89. Moreo, A., Romero, M., Castro, J. L., & Zurita, J. M. (2012). Lexicon-based comments-oriented news sentiment analyzer system. Expert Systems with Applications, 39(10): 9166-9180.

90. MorphoRuEval. (2017). Retrieved from https://github.com/dialogue-evaluation/morphoRuEval-2017/tree/social_media/.

91. Mosteller, J., Mathwick, C. (2016). Online Engagement Reviewer Journal of Service Research, 20 (2): 204-218.

92. Mount, J., Zumel, N. (2016). Exploring Data Science. Manning Publications.

93. Nakov, P. (2017). Semantic Sentiment Analysis of Twitter Data. arXiv: 1-15.

94. Nassirtoussi, A., Aghabozorgi, S., Wah, T., Ngo, D. (2014). Text mining for market prediction: A systematic review. Expert Systems with Applications, vol. 41, issue 16: 7653-7670.

95. Nlpub. (2017). Retrieved from https://nlpub.ru/Russian_Distributional_Thesaurus/.

96. O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM '10: 122-129.

97. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In: LREC, vol. 10.

98. Pang, B., Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '05: 115-124.

99. Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '02: 79-86.

100. Poecze, F., Ebster, C., & Strauss, C. (2018). Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. Procedia Computer Science, 130(C): 660-666.

101. Polyakov, P., Kalinina, M., Pleshko, V. (2012). Research on applicability of thematic classification methods to the problem of book review classification. In: Papers from the Annual International Conference 'Dialogue', vol. 11(2): 51-59.

102. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jimenez-Zafra, S.M., Eryigit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16: 19-30.

103. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I. (2015). SemEval-2015 task 12: Aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15: 486-495.

104. Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14: 27-35.

105. Pozzi, F., Fersini, E. (2017). Sentiment Analysis in Social Networks. Elsevier.

106. Provost, F., & Fawcett, T. (2013). Data science for business. Sebastopol, CA, O'Reilly.

107. PyPi. (2018). Retrieved from https://pypi.org/project/python-twitter/.

108. Qiu, J., Liu, C., Li, Y., & Lin, Z. (2018). Leveraging sentiment analysis at the aspects level to predict ratings of reviews. Information Sciences, 451: 295-309.

109. Qiu, L., Rui, H., & Whinston, A. (2013). Social network-embedded prediction markets: The effects of information acquisition and communication on predictions. Decision Support Systems, 55(4): 978-987.

110. Rambocas, M., Pacheco, B.(2018). Online sentiment analysis in marketing research: a review. Journal of Research in Interactive Marketing.

111. Rapp, A., Beitelspacher, L., Grewal, D., Hughes, D. (2013). Understanding social media effects across seller, retailer, and consumer interactions. Journal of the Academy of Marketing Science, vol. 41, issue 5: 547-566.

112. Repustate. (2018). Retrieved from https://www.repustate.com/.

113. ROMIP. (2010-2015). Retrieved from http://romip.ru/.

114. Rubtsova, Y. (2013). A method for development and analysis of short text corpus for the review classification task. In: Proceedings of Conferences Digital Libraries: Advanced Methods and Technologies, Digital Collections, RCDL: 269-275.

115. Rubtsova, Y. (2015). Constructing a corpus for sentiment classification training. “Programmnye produkty i sistemy” (Software & Systems), vol. 1 (109): 72-78.

116. Rubtsova, Y. (2017). Reducing the Degradation of Sentiment Analysis for Text Collections Spread over a Period of Time. Knowledge Engineering and Semantic Web: 8th International Conference: 3-13.

117. Rui, H., Liu, Y., & Whinston, A. (2013). Whose and what chatter matters? The effect of tweets on movie sales. Decision Support Systems, 55(4): 863-870.

118. RuSentiLex. (2017). Retrieved from http://www.labinform.ru/pub/rusentilex/index.htm/.

119. Russian National Corpus. (2018). Retrieved from http://www.ruscorpora.ru/.

120. RUSSIR. (2016-2017). Retrieved from http://romip.ru/russir/.

121. Russo, I., Caselli, T., Strapparava, C. (2015). SemEval-2015 task 9: CLIPEval implicit polarity of events. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval ’15: 442-449.

122. Rusvectores. (2018). Retrieved from http://rusvectores.org/ru/models/.

123. Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. Information Processing & Management, 52(1): 5-19.

124. Sapountzi A., Psannis Kostas E. (2016). Social networking data analysis tools & challenges. Future Generation Computer Systems.

125. Schutt, R., & O'Neil, C. (2014). Doing data science. Sebastopol, CA: O'Reilly Media.

126. scikit-learn. (2018). Retrieved from http://scikit-learn.org/.

127. Selegey, D., Shavrina, T., Selegey, V., & Sharoff, S. (2016). Automatic morphological tagging of russian social media corpora: training and testing. In Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue.

128. Semina, T. (2018). Subjectivity vs. Objectivity dichotomy and sentiment relevance in sentiment analysis tasks. Bulletin of the Moscow Region State University. Series: Linguistics, vol. 1: 36-48.

129. Sheng, J., Amankwah-Amoah, J., Wang, X. (2017). Multidisciplinary perspective of big data in management research. International Journal of Production Economics, vol. 191: 97-112.

130. Shriver, S., Nair, H., Hofstetter, R. (2013). Social Ties and User-Generated Content: Evidence from an Online Social Network. Management Science, vol. 59, issue 6: 1425-1443.

131. Smirnova, O. S., & Shishkov, V. V. (2016). The choice of the topology of neural networks and their use for the classification of small texts. International Journal of Open Information Technologies, 4(8): 50-54.

132. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4): 427-437.

133. Song, B., Ni, Y., Ren, Y., & Li, R. (2016). Business Trends in the Digital Era: Evolution of Theories and Applications.

134. Sonnier, G. P., McAlister, L., & Rutz, O. J. (2011). A dynamic model of the effect of online communications on firm sales. Marketing Science, 30(4): 702-716.

135. Sorokin, A. Shavrina, T., Lyashevskaya, O., Bocharov, V., Alexeeva, S., Droganova, K., Fenogenova, A., Granovsky, D. (2017). MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017”.

136. Statista. (2017). Retrieved from https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

137. Statista. (2018). Retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

138. Steinberger, J., Lenkova, P., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Steinberger, R., Tanev, H., Zavarella, V., Vazquez, S. (2011). Creating sentiment dictionaries via triangulation. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011: 28-36.

139. Stoyanov, V., Cardie, C. (2008). Topic identification for fine-grained opinion analysis. In: Proceedings of the 22nd International Conference on Computational Linguistics, COLING '08: 817-824.

140. Tadviser, (2017). Overview of Big Data in Russia. Retrieved from www.tadviser.ru/index.php/Статья:Большие_данные_(Big_Data)_в_России.

141. Tellez, E., Jimenez, S., Graff, M., Moctezuma, D., Suarez, R., Siordia, O. (2017). A simple approach to multilingual polarity classification in Twitter. Pattern Recognition Letters, vol. 94: 68-74.

142. Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. Marketing Science, 31(2): 198-215.

143. Tubishat, M., Idris, N., & Abushariah, M. A. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. Information Processing & Management, 54(4): 545-563.

144. Twitter. (2018). Retrieved from https://apps.twitter.com/.

145. Vasilyev, V., Denisenko, A., Solovyev, D. (2016). Aspect Extraction and Twitter Sentiment Classification by Fragment Rules. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016".

146. Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. Information Processing & Management, 53(3): 595-607.

147. Wiebe, J., Mihalcea, R. (2006). Word sense and subjectivity. In: ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics: 1065-1072.

148. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M. (2004). Learning subjective language. Comput. Linguist. 30(3): 277-308.

149. Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39(2-3): 165-210.

150. Wilson, T., Wiebe, J., Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT-EMNLP '05: 347-354.

151. Yablonsky, S. (2003). Russian morphology: resources and Java software applications.

152. Yadav M., Rahman, Z. (2017). Measuring consumer perception of social media marketing activities in e-commerce industry: Scale development & validation. Telematics and Informatics, vol. 34, issue 7: 1294-1307.

153. Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for Natural Language Processing. arXiv preprint arXiv:1702.01923.

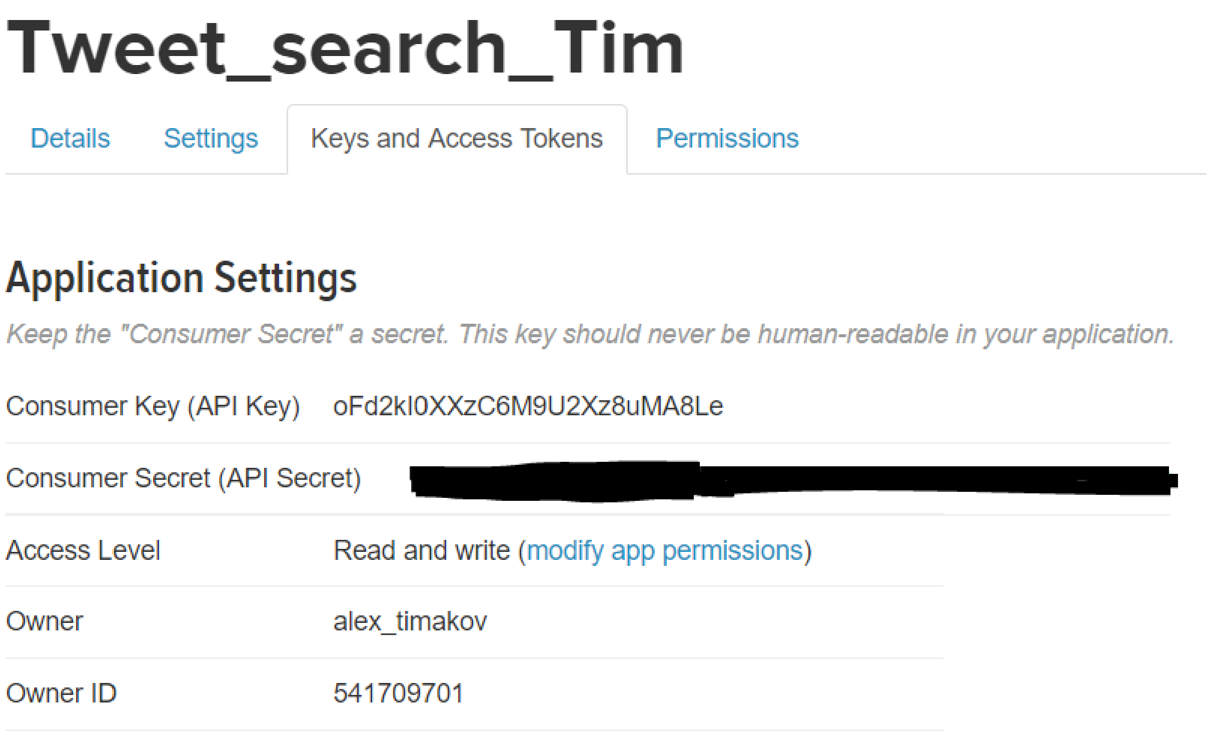
154. Yoon, J., Hwang, B. (2013). Design of a Productivity Measuring System for Twitter Marketing Reflecting User's Sentiment in Social Networks, Journal of KIISE, 40(6), 377-385.

155. Yussupova, N., Bogdanova, D., Boyko, M. (2012). Decision support for quality management based on artificial intelligence applications for unstructured data analysis. 1-12.

156. Zagibalov, T., Belyatskaya, K., Carroll, J. (2010). Comparable english-russian book review corpora for sentiment analysis. Computational Approaches to Subjectivity and Sentiment Analysis: 67–72.

Appendices

Appendix 1. Getting access to Twitter data



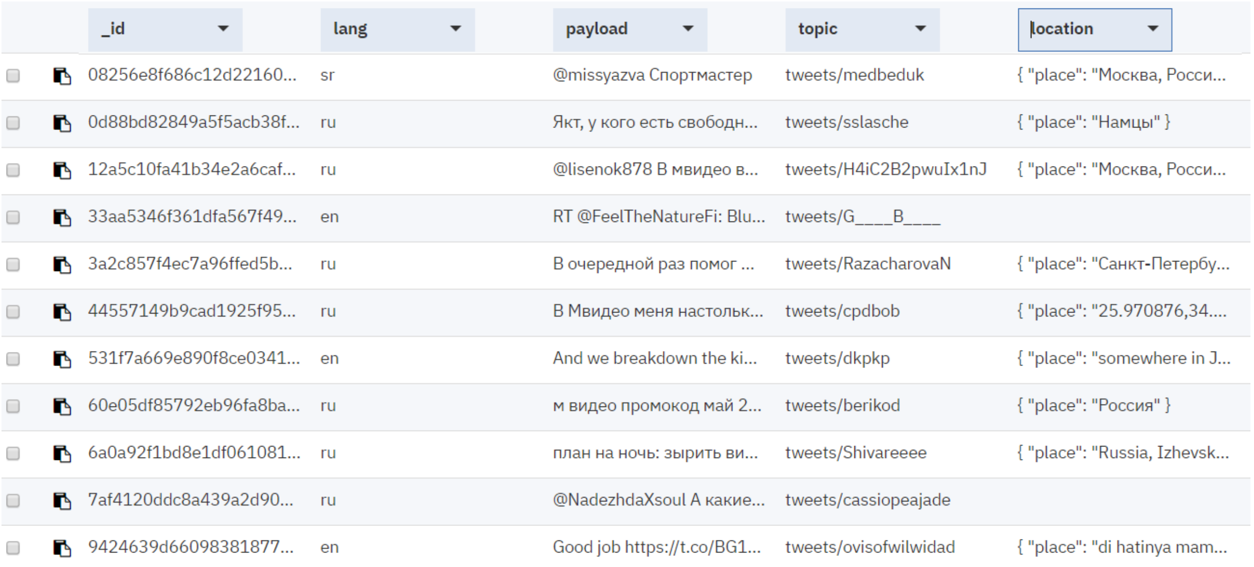
Tweet_search_Tim

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

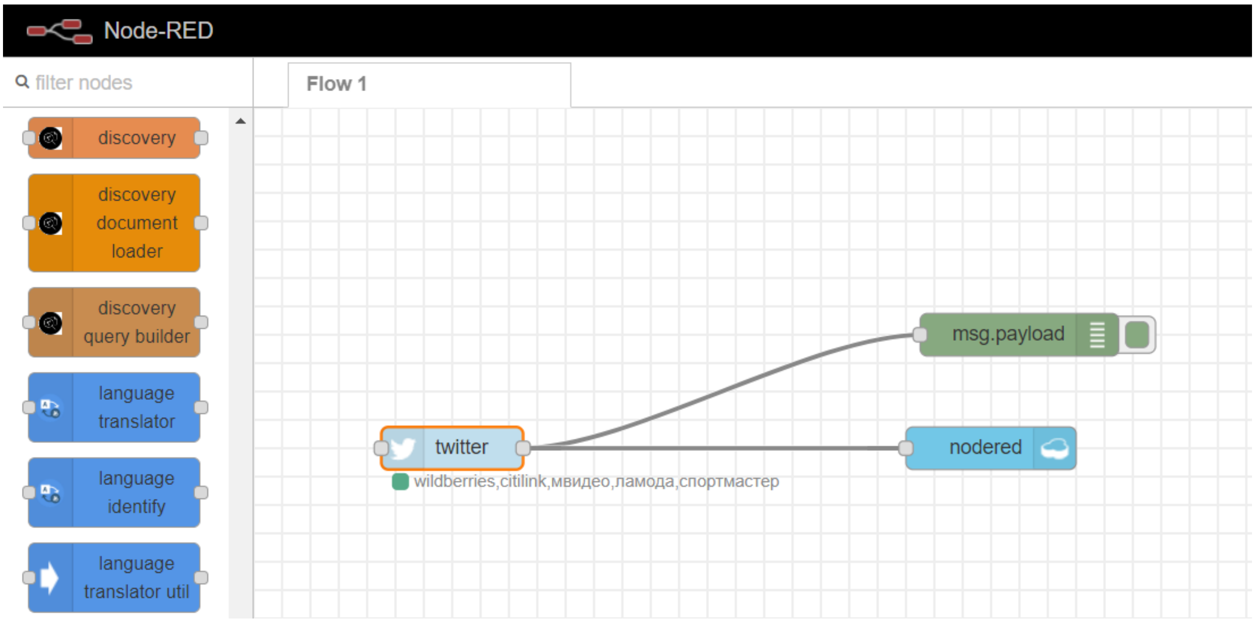
Consumer Key (API Key)	oFd2kI0XXzC6M9U2Xz8uMA8Le
Consumer Secret (API Secret)	[redacted]
Access Level	Read and write (modify app permissions)
Owner	alex_timakov
Owner ID	541709701

Appendix 2. NoSQL Cloudant database view

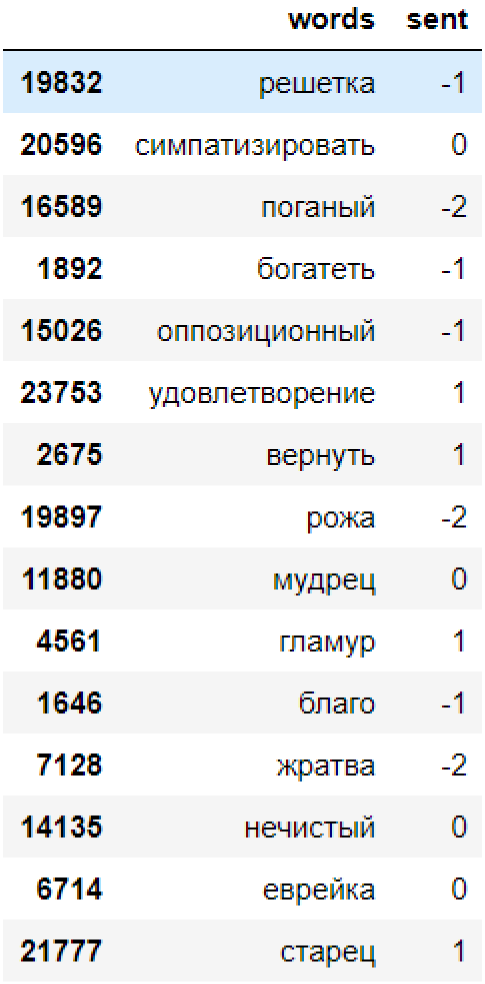


	_id	lang	payload	topic	location
□	08256e8f686c12d22160...	sr	@missyazva Спортмастер	tweets/medbeduk	{ "place": "Москва, Росси...
□	0d88bd82849a5f5acb38f...	ru	Якт, у кого есть свободн...	tweets/sslasche	{ "place": "Намцы" }
□	12a5c10fa41b34e2a6caf...	ru	@lisenok878 В мвидео в...	tweets/H4iC2B2pwuIx1nJ	{ "place": "Москва, Росси...
□	33aa5346f361dfa567f49...	en	RT @FeelTheNatureFi: Blu...	tweets/G____B____	
□	3a2c857f4ec7a96ffed5b...	ru	В очередной раз помог ...	tweets/RazacharovaN	{ "place": "Санкт-Петербу...
□	44557149b9cad1925f95...	ru	В Мвидео меня настольк...	tweets/cpdbob	{ "place": "25.970876,34....
□	531f7a669e890f8ce0341...	en	And we breakdown the ki...	tweets/dkpkp	{ "place": "somewhere in J...
□	60e05df85792eb96fa8ba...	ru	м видео промокод май 2...	tweets/berikod	{ "place": "Россия" }
□	6a0a92f1bd8e1df061081...	ru	план на ночь: зырить ви...	tweets/Shivareeee	{ "place": "Russia, Izhevsk...
□	7af4120ddc8a439a2d90...	ru	@NadezhdaXsoul А какие...	tweets/cassiopeajade	
□	9424639d66098381877...	en	Good job https://t.co/BG1...	tweets/ovisofwilwidad	{ "place": "di hatinya mam...

Appendix 3. Node-Red data collection

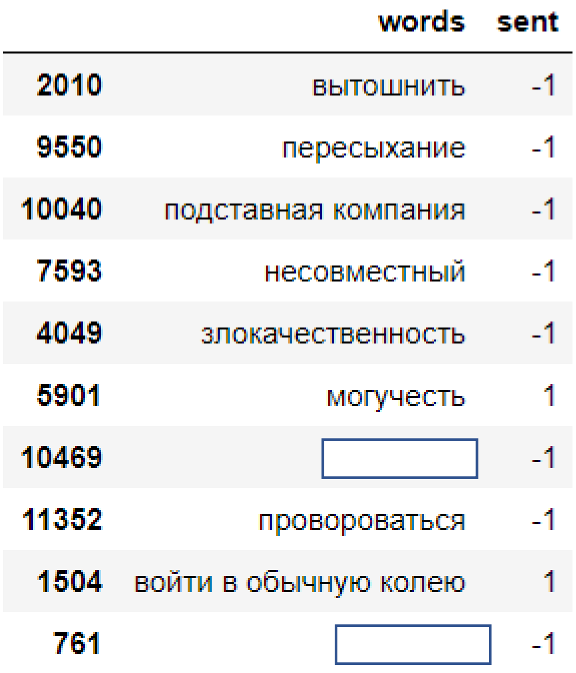


Appendix 4. Raw seed lexicon' example



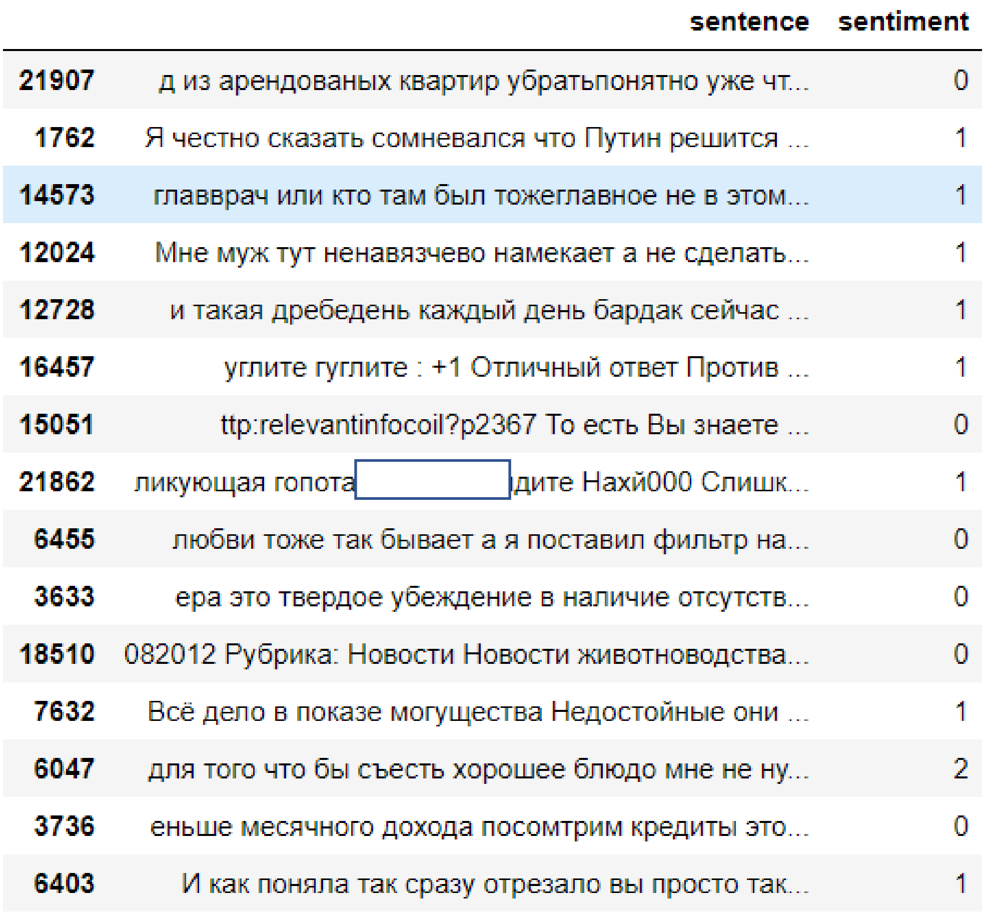
	words	sent
19832	решетка	-1
20596	симпатизировать	0
16589	поганый	-2
1892	богатеть	-1
15026	оппозиционный	-1
23753	удовлетворение	1
2675	вернуть	1
19897	рожа	-2
11880	мудрец	0
4561	гламур	1
1646	благо	-1
7128	жратва	-2
14135	нечистый	0
6714	еврейка	0
21777	старец	1

Appendix 5. Raw extended lexicon’ example



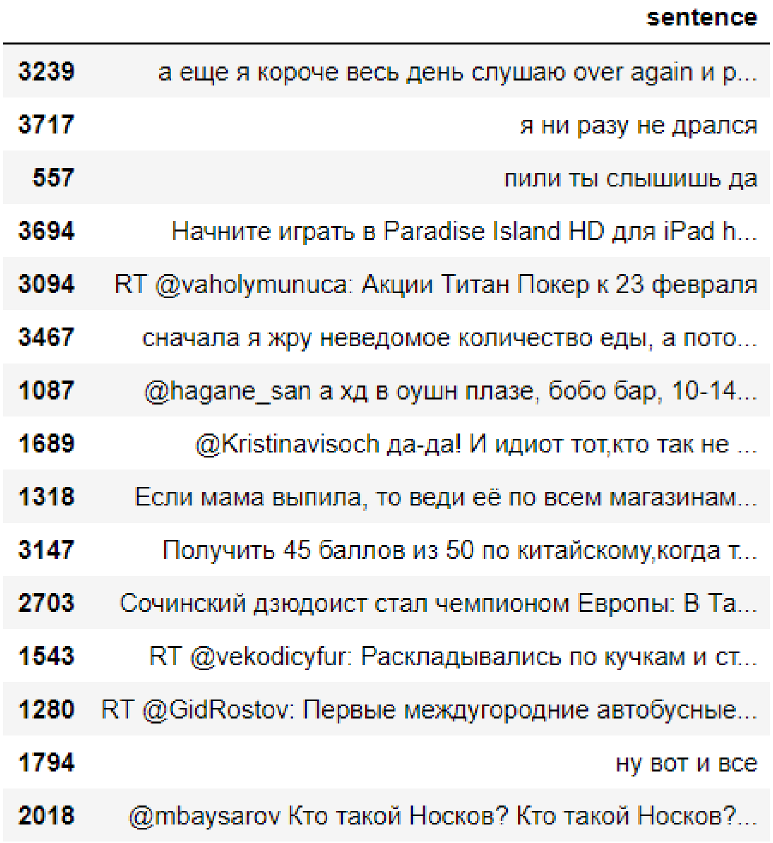
	words	sent
2010	вытошнить	-1
9550	пересыхание	-1
10040	подставная компания	-1
7593	несовместный	-1
4049	злокачественность	-1
5901	могучесть	1
10469	[redacted]	-1
11352	провороваться	-1
1504	войти в обычную колею	1
761	[redacted]	-1

Appendix 6. Raw labeled text corpora



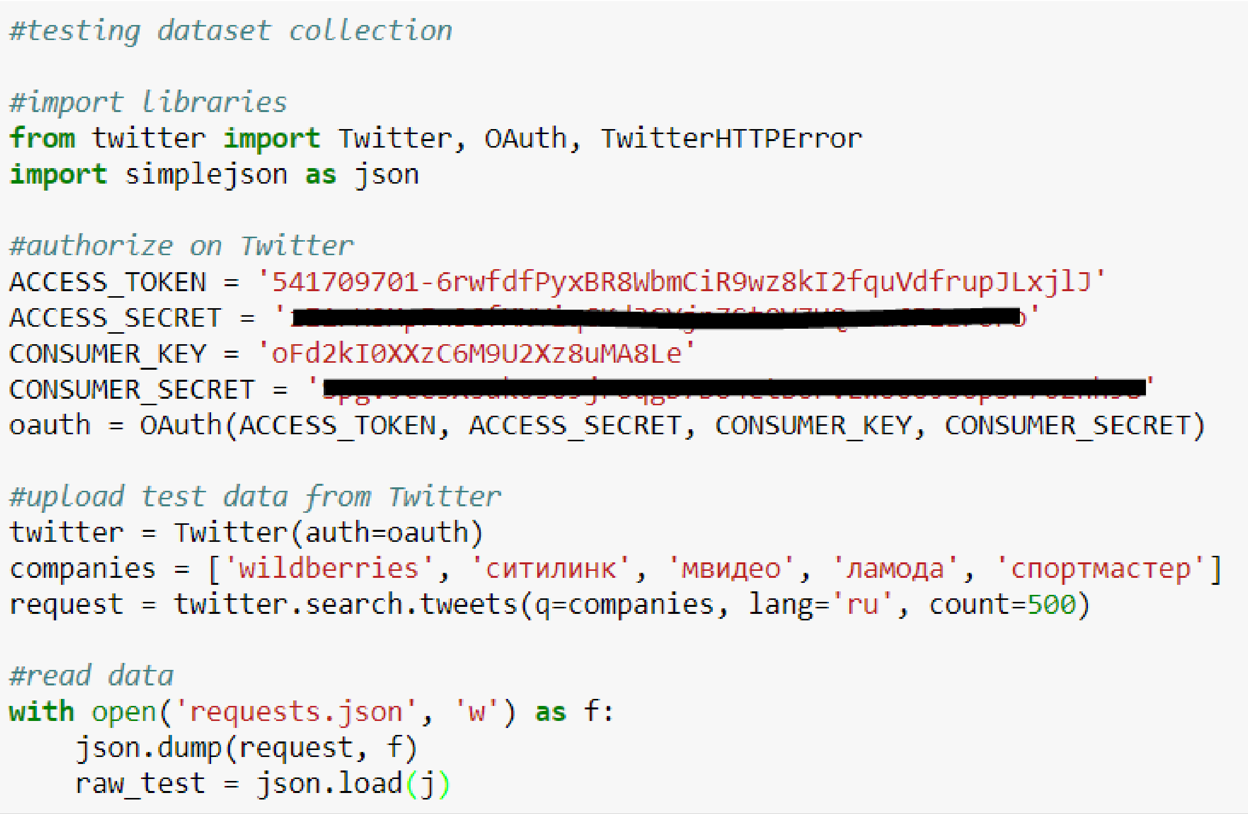
	sentence	sentiment
21907	д из арендованых квартир убратьпонятно уже чт...	0
1762	Я честно сказать сомневался что Путин решится ...	1
14573	главврач или кто там был тожеглавное не в этом...	1
12024	Мне муж тут ненавязчево намекает а не сделать...	1
12728	и такая дребедень каждый день бардак сейчас ...	1
16457	углите гуглите : +1 Отличный ответ Против ...	1
15051	ttp:relevantinfocoil?p2367 То есть Вы знаете ...	0
21862	ликующая гопота[redacted]дите Нахй000 Слишк...	1
6455	любви тоже так бывает а я поставил фильтр на...	0
3633	ера это твердое убеждение в наличие отсутств...	0
18510	082012 Рубрика: Новости Новости животноводства...	0
7632	Всё дело в показе могущества Недостойные они ...	1
6047	для того что бы съесть хорошее блюдо мне не ну...	2
3736	еньше месячного дохода посомтрим кредиты это...	0
6403	И как поняла так сразу отрезало вы просто так...	1

Appendix 7. Raw unlabeled text corpora



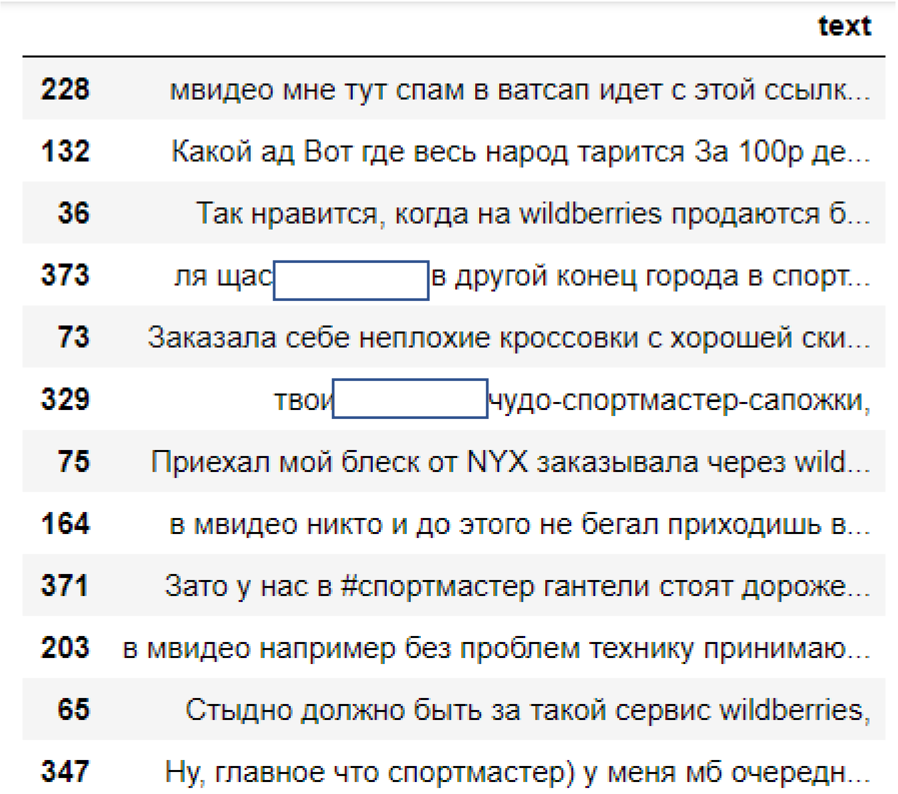
	sentence
3239	а еще я короче весь день слушаю over again и р...
3717	я ни разу не дрался
557	пили ты слышишь да
3694	Начните играть в Paradise Island HD для iPad h...
3094	RT @vaholymunuca: Акции Титан Покер к 23 февраля
3467	сначала я жру неведомое количество еды, а пото...
1087	@hagane_san а хд в оушн плазе, бобо бар, 10-14...
1689	@Kristinavisoch да-да! И идиот тот,кто так не ...
1318	Если мама выпила, то веди её по всем магазинам...
3147	Получить 45 баллов из 50 по китайскому,когда т...
2703	Сочинский дзюдоист стал чемпионом Европы: В Та...
1543	RT @vekodicyfur: Раскладывались по кучкам и ст...
1280	RT @GidRostov: Первые междугородние автобусные...
1794	ну вот и все
2018	@mbaysarov Кто такой Носков? Кто такой Носков?...

Appendix 8. Code for Testing tweets dataset collection



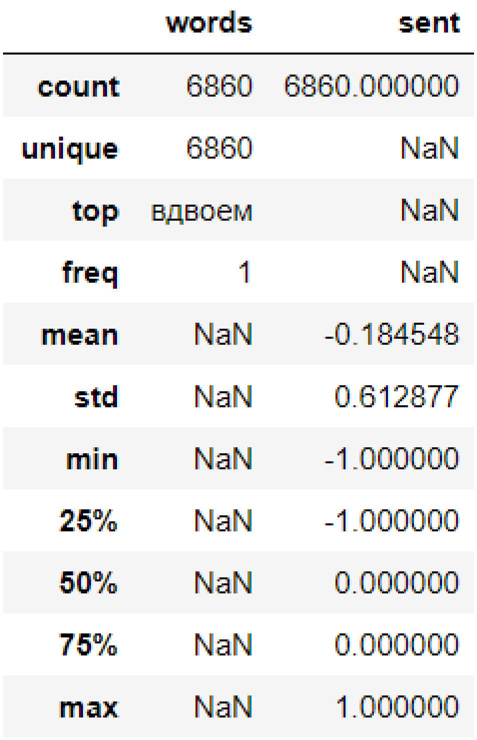
```
#testing dataset collection

#import libraries
from twitter import Twitter, OAuth, TwitterHTTPError
import simplejson as json

#authorize on Twitter
ACCESS_TOKEN = '541709701-6rwfdfPyxBR8WbmCiR9wz8kI2fquVdfrupJLxjlJ'
ACCESS_SECRET = '[redacted]'
CONSUMER_KEY = 'oFd2kI0XXzC6M9U2Xz8uMA8Le'
CONSUMER_SECRET = '[redacted]'
oauth = OAuth(ACCESS_TOKEN, ACCESS_SECRET, CONSUMER_KEY, CONSUMER_SECRET)

#upload test data from Twitter
twitter = Twitter(auth=oauth)
companies = ['wildberries', 'ситилинк', 'мвидео', 'ламода', 'спортмастер']
request = twitter.search.tweets(q=companies, lang='ru', count=500)

#read data
with open('requests.json', 'w') as f:
    json.dump(request, f)
    raw_test = json.load(j)
```

Appendix 9. Raw testing tweet sample

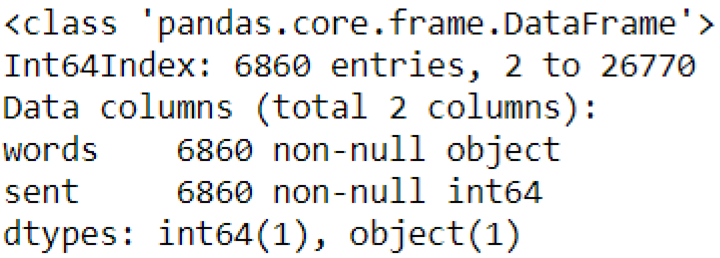


	text
228	мвидео мне тут спам в ватсап идет с этой ссылк...
132	Какой ад Вот где весь народ тарится За 100р де...
36	Так нравится, когда на wildberries продаются б...
373	ля щас в другой конец города в спорт...
73	Заказала себе неплохие кроссовки с хорошей ски...
329	твои чудо-спортмастер-сапожки,
75	Приехал мой блеск от NYX заказывала через wild...
164	в мвидео никто и до этого не бегал приходишь в...
371	Зато у нас в #спортмастер гантели стоят дороже...
203	в мвидео например без проблем технику принимаю...
65	Стыдно должно быть за такой сервис wildberries,
347	Ну, главное что спортмастер) у меня мб очередн...

Appendix 10. Prepared seed lexicon description

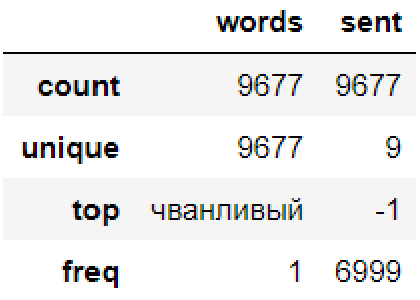


	words	sent
count	6860	6860.000000
unique	6860	NaN
top	вдвоем	NaN
freq	1	NaN
mean	NaN	-0.184548
std	NaN	0.612877
min	NaN	-1.000000
25%	NaN	-1.000000
50%	NaN	0.000000
75%	NaN	0.000000
max	NaN	1.000000

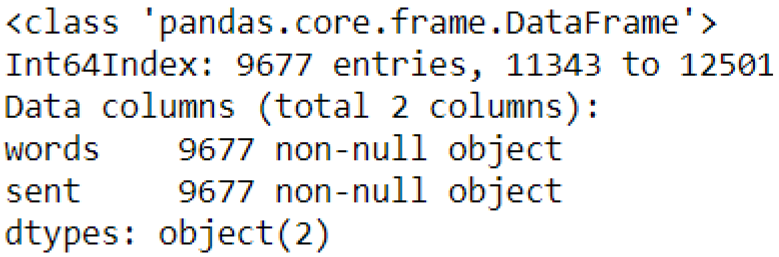


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6860 entries, 2 to 26770
Data columns (total 2 columns):
words    6860 non-null object
sent     6860 non-null int64
dtypes: int64(1), object(1)
```

Appendix 11. Prepared extended lexicon’ description

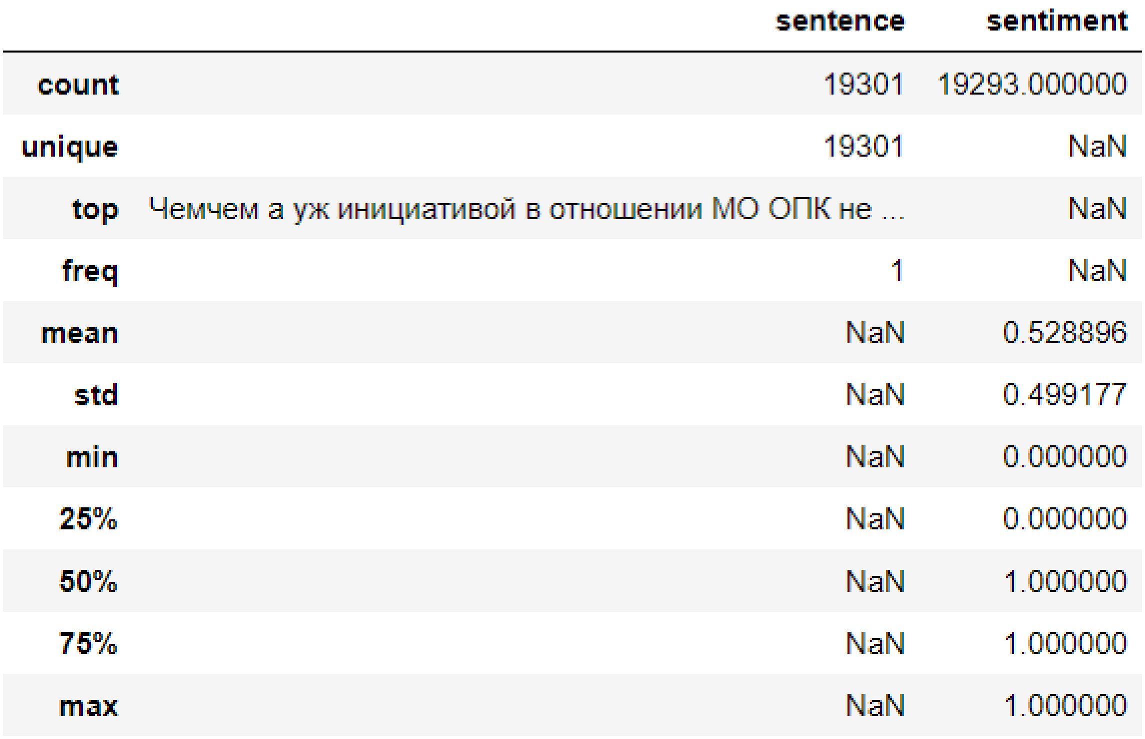


	words	sent
count	9677	9677
unique	9677	9
top	чванливый	-1
freq	1	6999

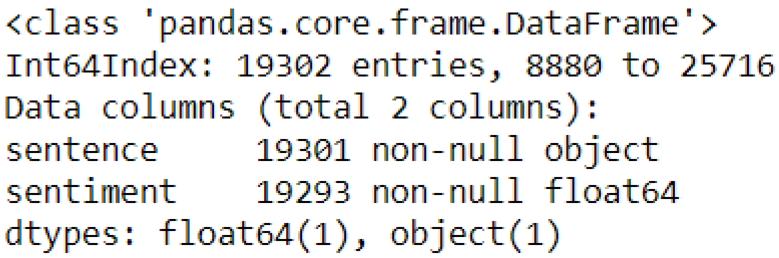


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9677 entries, 11343 to 12501
Data columns (total 2 columns):
words     9677 non-null object
sent      9677 non-null object
dtypes: object(2)
```

Appendix 12. Prepared training dataset for supervised learning’ description

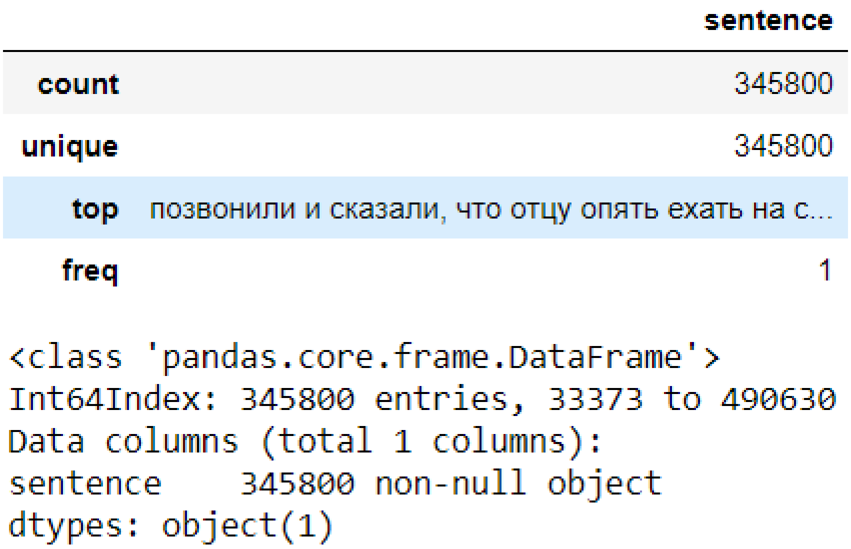


	sentence	sentiment
count	19301	19293.000000
unique	19301	NaN
top	Чемчем а уж инициативой в отношении МО ОПК не ...	NaN
freq	1	NaN
mean	NaN	0.528896
std	NaN	0.499177
min	NaN	0.000000
25%	NaN	0.000000
50%	NaN	1.000000
75%	NaN	1.000000
max	NaN	1.000000



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19302 entries, 8880 to 25716
Data columns (total 2 columns):
sentence     19301 non-null object
sentiment    19293 non-null float64
dtypes: float64(1), object(1)
```

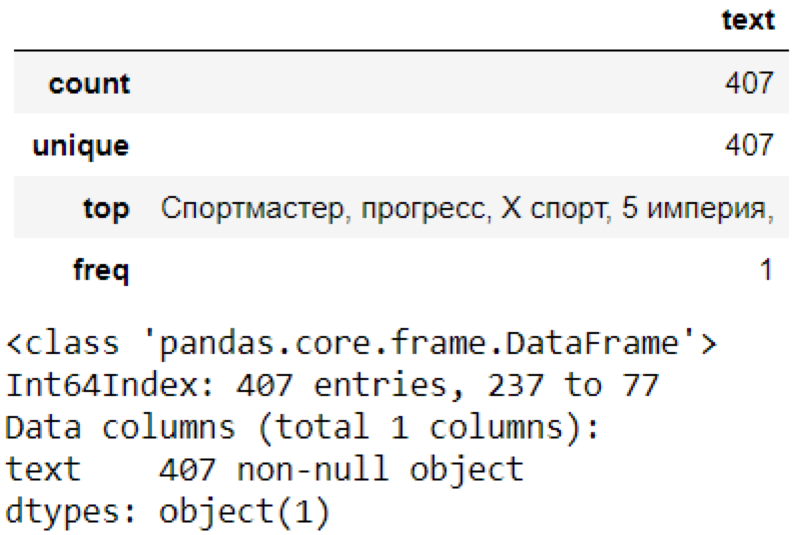
Appendix 13. Prepared training dataset for unsupervised learning' description



	sentence
count	345800
unique	345800
top	позвонили и сказали, что отцу опять ехать на с...
freq	1

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 345800 entries, 33373 to 490630
Data columns (total 1 columns):
sentence    345800 non-null object
dtypes: object(1)
```

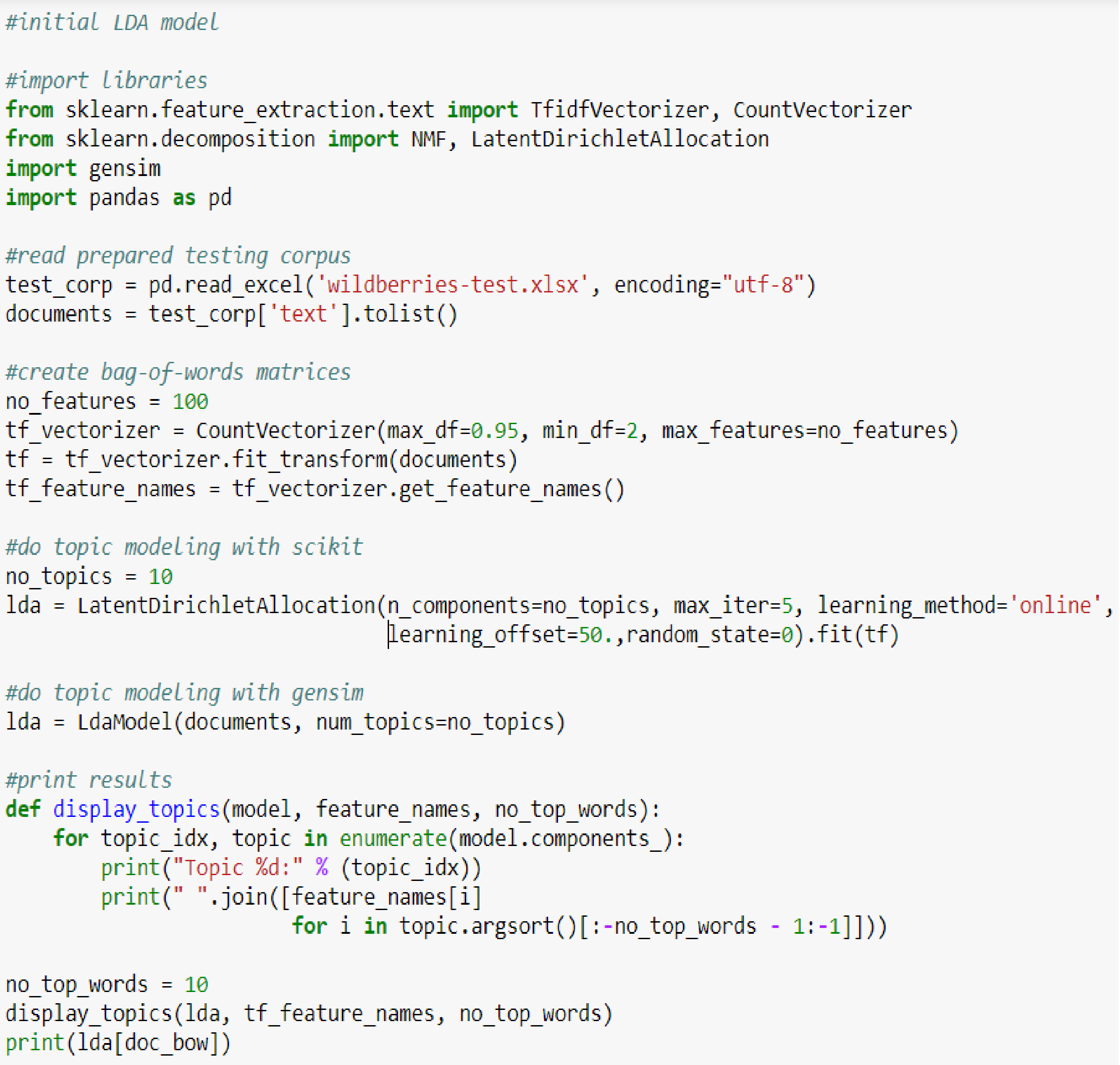
Appendix 14. Prepared testing tweet description



	text
count	407
unique	407
top	Спортмастер, прогресс, X спорт, 5 империя,
freq	1

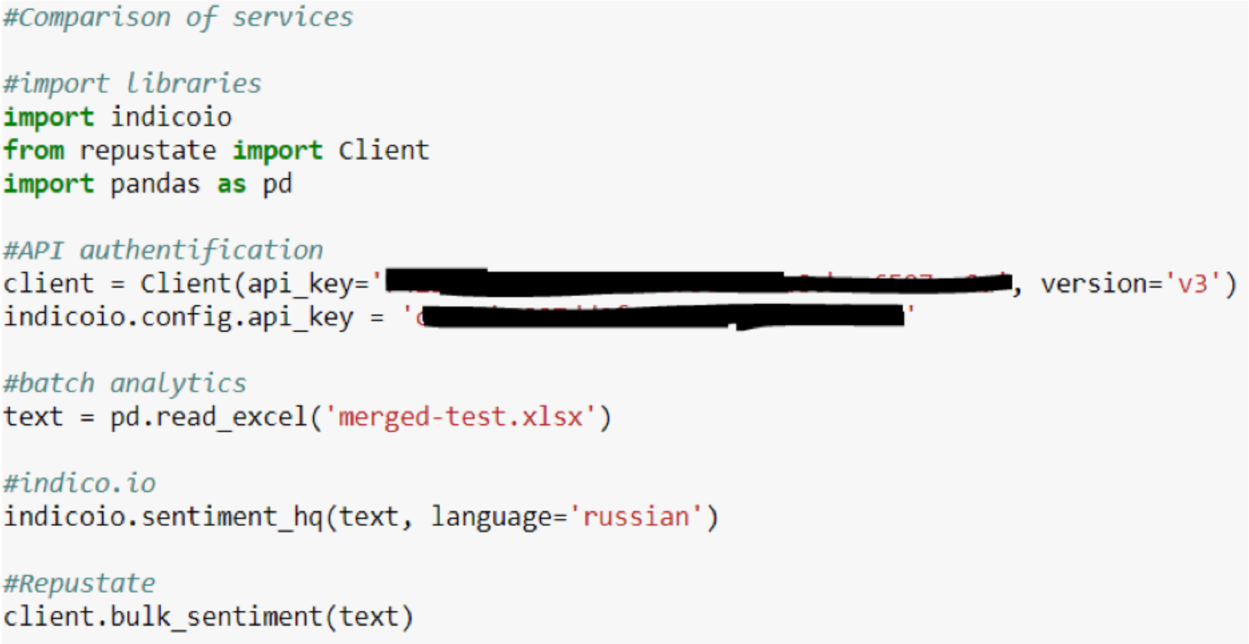
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 407 entries, 237 to 77
Data columns (total 1 columns):
text    407 non-null object
dtypes: object(1)
```

Appendix 15. Code for Initial LDA model



```python
#initial LDA model

#import libraries
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.decomposition import NMF, LatentDirichletAllocation
import gensim
import pandas as pd

#read prepared testing corpus
test_corp = pd.read_excel('wildberries-test.xlsx', encoding="utf-8")
documents = test_corp['text'].tolist()

#create bag-of-words matrices
no_features = 100
tf_vectorizer = CountVectorizer(max_df=0.95, min_df=2, max_features=no_features)
tf = tf_vectorizer.fit_transform(documents)
tf_feature_names = tf_vectorizer.get_feature_names()

#do topic modeling with scikit
no_topics = 10
lda = LatentDirichletAllocation(n_components=no_topics, max_iter=5, learning_method='online',
                                learning_offset=50.,random_state=0).fit(tf)

#do topic modeling with gensim
lda = LdaModel(documents, num_topics=no_topics)

#print results
def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print("Topic %d:" % (topic_idx))
        print(" ".join([feature_names[i]
                        for i in topic.argsort()[:-no_top_words - 1:-1]]))

no_top_words = 10
display_topics(lda, tf_feature_names, no_top_words)
print(lda[doc_bow])
```

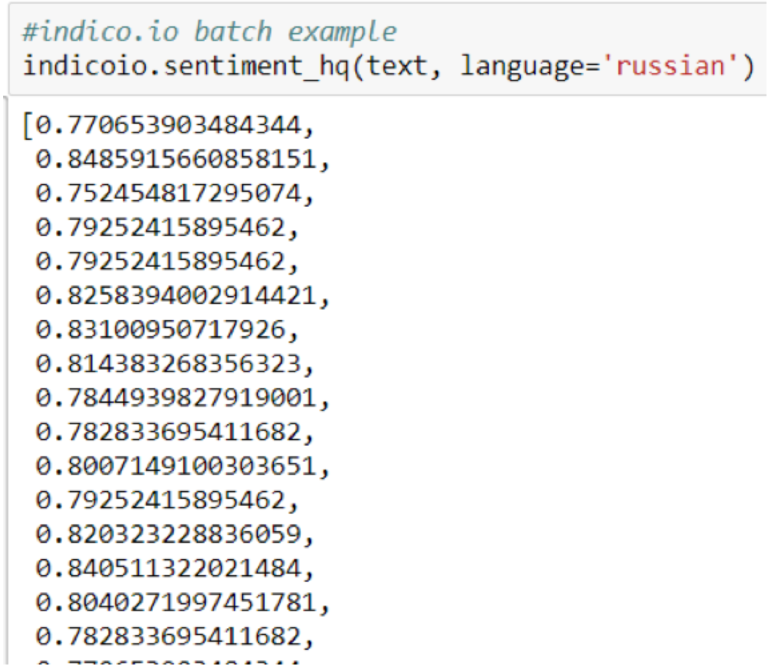
Appendix 16. Code for Comparison of initial models with existing services



```
#Comparison of services

#import libraries
import indicoio
from repustate import Client
import pandas as pd

#API authentification
client = Client(api_key='[redacted], version='v3')
indicoio.config.api_key = '[redacted]'

#batch analytics
text = pd.read_excel('merged-test.xlsx')

#indico.io
indicoio.sentiment_hq(text, language='russian')

#Repustate
client.bulk_sentiment(text)
```

Appendix 17. Polarity classification output by existing services

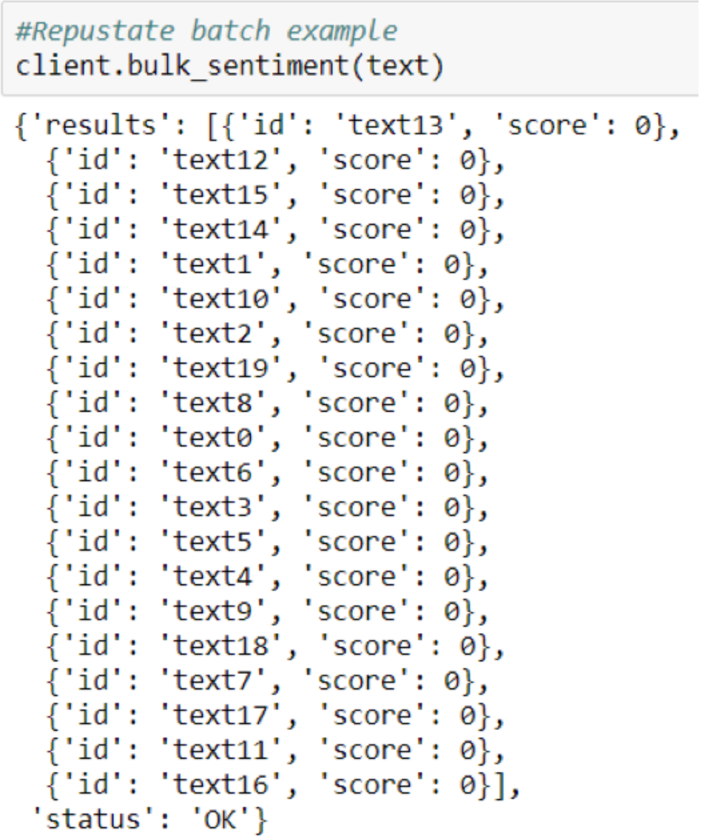
indico.io



```
#indico.io batch example
indicoio.sentiment_hq(text, language='russian')
```

```
[0.770653903484344,
 0.8485915660858151,
 0.752454817295074,
 0.79252415895462,
 0.79252415895462,
 0.8258394002914421,
 0.83100950717926,
 0.814383268356323,
 0.7844939827919001,
 0.782833695411682,
 0.8007149100303651,
 0.79252415895462,
 0.82032322883605 9,
 0.840511322021484,
 0.8040271997451781,
 0.782833695411682,
```

Repustate



```
#Repustate batch example
client.bulk_sentiment(text)
```

```
{'results': [{'id': 'text13', 'score': 0},
  {'id': 'text12', 'score': 0},
  {'id': 'text15', 'score': 0},
  {'id': 'text14', 'score': 0},
  {'id': 'text1', 'score': 0},
  {'id': 'text10', 'score': 0},
  {'id': 'text2', 'score': 0},
  {'id': 'text19', 'score': 0},
  {'id': 'text8', 'score': 0},
  {'id': 'text0', 'score': 0},
  {'id': 'text6', 'score': 0},
  {'id': 'text3', 'score': 0},
  {'id': 'text5', 'score': 0},
  {'id': 'text4', 'score': 0},
  {'id': 'text9', 'score': 0},
  {'id': 'text18', 'score': 0},
  {'id': 'text7', 'score': 0},
  {'id': 'text17', 'score': 0},
  {'id': 'text11', 'score': 0},
  {'id': 'text16', 'score': 0}],
 'status': 'OK'}
```