СТОПАНСКИ ФАКУЛТЕТ
SOFIA UNIVERSITY ST. KLIMENT OHRIDSKI est. 1888
FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

Department of "Statistics and Econometrics"

GLORIA VENTSISLAVOVA HRISTOVA

AN AUTOMATED SYSTEM FOR ANALYSIS OF ONLINE COMMUNICATION WITH CUSTOMERS BASED ON MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING - STRUCTURE, DEVELOPMENT AND BUSINESS APPLICATIONS

DISSERTATION ABSTRACT

for the award of
educational and scientific degree "Doctor"
Professional field: 3.8. Economics,
Scientific specialty: Data Science

Supervisor: **Assoc. Prof. Boryana Bogdanova**

Sofia, 2022

The dissertation is approved and directed for public defense by the council meeting of the Department of Statistics and Econometrics at the Faculty of Economics and Business Administration at Sofia University “St. Kliment Ohridski”, held on June 20, 2022 (Protocol № 331).

The author of the dissertation is a full-time PhD student at the Department of Statistics and Econometrics, according to a decision of the Faculty Council at the Faculty of Economics and Business Administration, order № RD 20-321/05.02.2019 of the Rector of Sofia University.

The dissertation consists of an introduction, main text (three chapters), conclusion, bibliography and appendices with a total volume of 281 pages, 37 figures and 31 tables.

The cited literature includes 220 sources of Bulgarian and foreign authors.

Number of publications related to the dissertation: 5.

Contents

I. General characteristics of the dissertation 4

1.1. Relevance of the research topic 4

1.2. Object and subject of the study 7

1.3. Main aim and research objectives 8

1.4. Research thesis and hypotheses 9

1.5. Research method 11

1.6. Scope of the study 11

1.7. Structure of the dissertation 12

II. Brief description of the dissertation 13

2.1. Literature review (Chapter I) 13

2.1.1. Chat data analysis 13

2.1.2. Text analytics in Bulgarian 15

2.1.3. Topic modeling and sentiment analysis applied on text data 19

2.2. Methodology (Chapter II) 22

2.2.1. Module I 23

2.2.2. Module II 24

2.2.3. Module III 26

2.2.4. Module IV 29

2.3. Empirical study (Chapter III) 30

2.3.1. Topic modeling (Module II) 31

2.3.2. Customer sentiment analysis (Module III) 36

2.3.3. Summary and visualization of results (Module IV) 41

III. Conclusion 43

IV. Bibliography 45

V. Contributions of the dissertation 47

VI. List of publications related to the dissertation 48

I. General characteristics of the dissertation[1]

1.1. Relevance of the research topic

In the XXI[st] century, the concept of **"big data"** has emerged. This sparked the interest of business companies, set new horizons and opened the door to "data science". According to Statista's forecast, the amount of data generated worldwide will increase to over 180 zetabytes by 2025[2]. **Data acquisition and implementation of algorithmic solutions** with the help of machine learning are no longer a competitive advantage, but a necessity for companies in order to remain competitive. The business is in a **period of transformation**, the driving force being data science and machine learning, which allow: **process automation; drawing valuable insights into customer behavior; cost cutting; products and services improvement** and much more.

The increasing competition leads to placing customers and their satisfaction at a central place in modern business models [1] and data becomes a valuable asset for companies - a source of information on the **important topics of interest for customers, their preferences, needs, problems, sentiments and overall satisfaction with products and services** [2]. According to a study conducted by Dimension Data, which includes 1 351 industry organizations in 80 countries [3] - **84%** of companies aimed at **increasing customer satisfaction** report an increase in revenue as a result of their efforts.

A successful example of an **approach in which the clients' topics of interest, as well as their needs and preferences are central**, is that of The New York Times. An article published in January 2021[3] talks about the company's experience in implementing solutions based on machine learning in order to get closer to its main customer - the reader. The media giant is applying topic modeling algorithms to analyze the topics of customer interest, with the aim of articles recommendation that best meets customers' interests.

[1] The dissertation is developed as part of the first Industry-funded PhD program at the Faculty of Economics and Business Administration, Sofia University. In the recent years, the business has realized the need to develop and integrate business solutions based on funded scientific research. In this regard, the dissertation's introduction contains not only the classical attributes and framework of the research but also motivates in detail the need to develop scientific research aimed at solving current business problems.

[2] Statista's forecasts for the growth in the volume of data generated globally are available at the following address: https://www.statista.com/statistics/871513/worldwide-data-created/

[3] The article is available in digital format at the following address - https://open.nytimes.com/we-recommend-articles-with-a-little-help-from-our-friends-machine-learning-and-reader-input-e17e85d6cf04

Companies from different industries try to reach their customers in all possible ways in order to better understand their needs and respond to them as efficiently and quickly as possible [2]. In this regard, **data from communication** with the client in various forms - **interviews, customer reviews, data from social networks** and others is beginning to play a central role [2], [4]. In 2020, another factor emerged in this complex picture of dynamic relations between the company and the client – namely, the COVID-19 pandemic. The last has become a catalyst for an even greater business transformation and digitalization of many processes and services. One of the most noticeable transformations is in terms of **communication with the client**.

Many end-user companies face the challenges of **fully digital customer communication**. Opportunities for physical meetings with clients and surveys of their satisfaction with services through classical approaches (in-depth interviews and focus groups) are severely limited due to the epidemic situation [5]. **There is a need** to discover new approaches to the analysis of customers' behavior, satisfaction, topics of interest and needs, based entirely on the online communication with them.

Among the affected industries worldwide, **the banking sector is undergoing a huge transformation**. A KPMG report [6] states that as a result of the pandemic, banks have been forced to digitize many processes and services much faster than planned, meanwhile trying to stay "close to the customer". The competition pressure in the face of the fintech industry exacerbates the problem. At the same time, the volume of customer data available to banks is increasing significantly - the forecasts in the banking sector for the period 2021-2026 are for a combined annual growth rate in the use of big data analysis technologies of about 22.97% [7].

Various studies before and after the COVID-19 pandemic clearly point the **trend to the digitalization of customer communication**. In 2020, compared to the previous year, the number of companies using the WhatsApp chat platform for customer service increased by as much as **62%** and the use of Facebook messenger for such purposes increased by **51%**[4]. In this boom of online communication in a variety of industries, **contact centers are at the forefront** - a huge part of the communication with the customer is conducted there. The role of contact centers as a **direct connection between the client and the company**, gives them an increasingly important and central place.

[4] The results of the study are available at the following address: https://www.statista.com/statistics/1260555/top-messaging-channels-for-customer-service-by-yoy-growth/

One of the most important aspects for the business in such communication are the **topics** of client interest, as well as customer **satisfaction** with the contact center services [8], [9], [10], [11]. Such data is a **reflection of the main problems and topics** that interest customers. This knowledge could serve as an indicator of what is most important to clients, what problems they have encountered, in what are they most interested in, and hence what would affect their opinion and decision to continue using the company's products and services. In customer communication within the contact center, it is of **huge importance to monitor satisfaction** - whether the operator has succeeded in fulfilling the request and whether the customer was satisfied with the provided service [12], [9].

In this regard, the COVID-19 pandemic brought even more attention as it made "digital services" a topic of **great importance to the business**. During a global pandemic, when physical meetings are severely limited, the client is "far away" and **can no longer be reached in the well-known traditional ways**. The **digitalization of communication** has led to an increase in the amount of data generated in contact centers - as mentioned earlier, **a large increase has been observed in the use of chat systems** to communicate with the client. Such text data is highly unstructured, generated in real-time and might be of huge volume. This makes it impossible to manually review and summarize topics of interest to the client, as well as customers' problems, satisfaction or other important indicators of the relationship between the company and them.

However, **the current challenges in front of the business also create new opportunities** – the dissertation is focused on the application of **text analytics** and **machine learning to online chat communication with the client**. Such methods are characterized by speed, thus allowing big data analysis and automation of the analysis of current trends in customer opinion, needs and satisfaction with the products and services that a company provides [9], [13], [14]. The critical literature review, carried out in the dissertation reveals that little research has been focused on this area, although **in the last two years online chat communication with the client has become a particularly important topic for the industry.**

The need to establish a **reliable methodology for quantitative analysis of online chat communication with the client** is at the core of the dissertation. Apart from the COVID-19 pandemic, what further complicates and makes this problem even more critical is that in such quantitative analyses, the language of textual communication plays a very important role [15]. The existing natural language processing tools as well as various analytical systems **work mainly with**

textual data in English. Textual data in Bulgarian is not so commonly used in the empirical studies and scientific literature devoted on quantitative analysis and processing of text data. The critical literature review, analysis and synthesis in Chapter I of the dissertation reveal that **so far, no research has been published on the analysis of online chat communication with clients in Bulgarian.**

All that has been said so far emphasizes **the importance and need not only of establishing a reliable methodology for quantitative analysis of online chat communication with clients**, but also **the importance of this topic when it comes specifically to text data in Bulgarian**. In an era of global economic and business transformation, when a huge number of different types of text data are generated every day, their quantitative analysis is about to attract more and more interest due to the potential it brings for the industry [13]. These trends inevitably affect Bulgaria as well, imposing the need for a critical literature review of the **opportunities and the current level of development in the scientific field dealing with the analysis of textual data specifically in the Bulgarian language.**

1.2. Object and subject of the study

The object of the dissertation **is online chat communication between clients and employees of a large company in Bulgaria**. The data from this communication is generated in the contact center of the company, which operates in the banking sector. In areas such as banking, COVID-19 has led to some radical changes in processes, services and overall relationship with the customer. If until recently customers visited the bank's office to ask in person their questions, today many of them prefer to interact with the company online[5]. This led to a huge amount of communication data, generated directly in banks' contact centers and operators were at the forefront during the pandemic crisis.

The **subject** of the dissertation are the **main topics of client interest, as well as customer satisfaction with contact center chat services.** Topics discussed in the chat communication between clients and the company are a valuable source of information about customers' interests,

[5] In an S&P survey conducted in February and March 2021, 52% of respondents said they visit physical offices of banks less frequently after the start of the COVID-19 pandemic. More interesting observations and insights into customer behavior in the banking sector are available in an article at the following address - https://www.spglobal.com/marketintelligence/en/news-insights/research/pandemic-pushes-customers-out-of-branches-banks-ramp-up-closures

problems and needs. Meanwhile, in an era of digital communication, when customer service plays a central role, it is of great importance to monitor **customers' satisfaction with communication** [4] with the company.

1.3. Main aim and research objectives

Main aim of the dissertation:

Development of an automated system for the analysis of both the main topics of client interest, as well as customer satisfaction with the services provided in a contact center in which communication is in Bulgarian.

A set of different **analytical techniques in the field of natural language processing** (adapted to the specifics of the Bulgarian language) and **machine learning** is utilized to achieve the main aim. A very important aspect of the main goal is the **focus on the analysis of textual data in Bulgarian**. The literature review reveals that this is a relatively underdeveloped scientific field. **Currently there is no systematic literature review and synthesis of the achieved by researchers in the field**. The knowledge is rather scattered and fragmented and **within the dissertation is made an attempt to introduce structure in this scientific field.**

Based on the findings made as a result from the conducted critical literature review, analysis and synthesis in Chapter I of the dissertation the following **research objectives** are formulated:

I. Conducting a critical literature review, analysis and synthesis in the following three areas of research interest:
   - I.1. Methodical literature review with respect to the object of the dissertation - online chat communication with clients carried out in a contact center.
   - I.2. Extensive literature review of research focused on the analysis of textual data in Bulgarian - development in the field and practical applications.
   - I.3. Methodical literature review with respect to the subject of the dissertation – topic modeling and customer satisfaction/sentiment analysis.

II. Development of an automated system for analysis of online communication with customers based on machine learning and natural language processing. Focus is put on the **structure** and **development** of the system:

II.1. Development of a research method for structuring and processing of the studied type of data (**Module I** of the system).

II.2. Development of a research method for extraction of the main topics of client interest (**Module II**). Adaptation of specific techniques in the context of working with textual data in Bulgarian.

II.3. Development of a research method for the analysis of customer satisfaction with contact center chat services (**Module III**). Adaptation of specific techniques in the context of working with textual data in Bulgarian.

III. Empirical testing and practical application of the developed system on a sample of data part of a real business case. Focus is put on the **business applications** of the system:

III.1. Formulation of clear and specific recommendations for the possible ways in which the knowledge extracted from the studied type of data could be useful for the business (**Module IV**).

III.2. Outlining opportunities for improvement of the developed system and future perspectives.

1.4. Research thesis and hypotheses

One **fundamental question** plays a key role in formulating the thesis of the dissertation, namely: "**Are there efficient ways to analyze and extract valuable information from the studied type of data using methods in the field of text analytics and machine learning?**" Although the answer may seem quite obvious and straightforward at first, an important factor is the extent to which such knowledge could be extracted from the studied type of data. Chats are characterized by various peculiarities and are more difficult to process compared to other types of textual data. Finding an answer to this question will shed light on the extent to which the information extracted from this type of data **adds value and is valuable to the business**. Thus, the **research thesis** underlying the dissertation is defined as follows:

Online chat communication between customers and operators in a contact center is an unexplored rich source of information about customers' relationship with the company. This information could be efficiently extracted, structured and analyzed with the help of techniques in the field of natural language processing and machine learning in order to build an automated analysis system characterized by valuable applications in the business.

The development of the research thesis is carried out within the framework of the following research questions (formed on the basis of the literature review in Chapter I):

1. **Are there efficient methods in the field of machine learning and text analytics, which could be applied to the object of the dissertation in order to extract the main topics of client interest?**

   Formulated hypotheses in regard to this research question:

   Hypothesis 1: *The traditional topic modeling method - Latent Dirichlet Allocation (LDA), could be applied to the object of the dissertation, and this will lead to results that are a reliable basis for developing a working solution.*

   Hypothesis 2: *In the application of an algorithm for modeling the topics of client interest, optimal results will be achieved (in terms of clarity and quality of the obtained topics) by using a sample of only client utterances, compared to using the entire chat between the customer and the operator.*

To the best of the author's knowledge, **there exists only one study** [16] **similar to the current one in terms of research objectives and studied type of data**, namely - prediction of the **final customer service rating** (measuring customer satisfaction with contact center chat services). The researchers make an important conclusion - **the existing sentiment lexicons for text data in English are not applicable in the customer services domain and do not lead to satisfactory results.** The availability of such resources for the Bulgarian language is very limited. Findings in [16] lead to the emergence of an **unexplored niche in the scientific literature** and the formulation of the following research question:

2. **Could it be assumed that the textual dialogue (chat) between a client and an operator, contains enough signals that might be used to create a customer satisfaction prediction model, which would be a reliable basis for developing a working solution?**

The answer to this question also determines the opportunity to form an answer to one much more general question, namely: **Is it possible to predict the satisfaction of customers based mainly on their text dialogues (chats) with the operators**? Based on the findings of other researchers in the field published in [16], the answer to this question is sought in the dissertation through the application of alternative techniques based entirely on machine learning. Formulated are the following two hypotheses:

Hypothesis 3: *It is possible to create an automated machine learning model, that predicts customers' satisfaction only on the basis of their chat with an operator. The developed model will be characterized by an optimal performance compared to making a "naïve" prediction which is not based on machine learning.*

Hypothesis 4: *Optimal results in customer satisfaction prediction could be achieved by using a sample of customers' final utterances in the chat communication (compared to what could be achieved by using alternative levels of chat data representation).*

1.5. Research method

The **research method** applied in the dissertation relies entirely on **quantitative methods for textual data analysis**. Techniques in the field of natural language processing and machine learning are applied in order to achieve the main goal of the current study. Such innovative approaches have the potential to **efficiently complement conventional techniques for analyzing customer behavior and satisfaction** and to **increase the degree of rationality** in management decisions. These methods are used not only due to efficiency reasons but also in order to automate the analysis of customer behavior in support of making better management decisions.

1.6. Scope of the study

The empirical study carried out in Chapter III is limited to the analysis of online chat communication generated in the contact center of a particular company - a large bank operating in Bulgaria. **This limitation does not affect the accomplished in Chapter I and Chapter II**, namely - an in-depth review, analysis and synthesis of research in the field, as well as development of a methodology for the analysis of online chat communication in a contact center (in the Bulgarian language).

Although the empirical study is focused on a specific business case in the banking domain, the author's proposed **method for interpretation of the results** is applicable in all other industries where such type of data is generated as a result of the company's business processes. **This applicability and potential opportunity for extrapolation of the steps in the interpretation and analysis of such type of data in other industries outlines one of the contributions of the dissertation.**

In addition to the **practical business applications**, the **current study might be of interest to** the research community in the field of text data analysis aimed at optimization and improvement in customer relationship management. Researchers focused on the analysis of textual data in the Bulgarian language might also be interested in the current study since it provides an in-depth overview of the development, current opportunities and future directions in the field. Last but not least, the dissertation might be found valuable by the research community focused on the study of chat communication, as well as on the analysis of such communication conducted precisely in a contact center.

1.7. Structure of the dissertation

In Chapter I is conducted a detailed literature review, analysis and synthesis of research articles focused on the main object and subject of research in the dissertation. Some basic concepts and hypotheses are defined, as well as the chosen analytical approach (based on the work of other researchers in the field). Chapter I also outlines gaps in the existing literature that the current study aims to fill. **The scope of Chapter I is broad**, as an extensive literature review, analysis and synthesis are carried out, aiming to present an up-to-date picture of the development in the field of textual data analysis in Bulgarian. Chapter II presents in detail each step in the methodology followed in order to achieve the main aim of the dissertation - building an automated system for analysis of online communication with customers based on machine learning and natural language processing. In Chapter III, the developed system is tested empirically in the context of a real business case - valuable conclusions are drawn regarding the applicability of the system and the future directions for its improvement.

II. Brief description of the dissertation

2.1. Literature review (Chapter I)

In Chapter I, a literature review is conducted in the following areas of interest:

- ✓ Current research on the analysis and information extraction from online chat communication (with the help of machine learning and natural language processing techniques). Special attention is paid to research focused on the analysis of chat communication in contact centers (the customer service domain).
- ✓ Research focused on the analysis of textual data in Bulgarian - development in the field and practical applications (a comprehensive literature review).
- ✓ Main approaches, current trends and established techniques for sentiment analysis and topic modeling of textual data.

The main results from the critical literature review, analysis and synthesis are published in [17] and [18].

2.1.1. Chat data analysis

The literature review of current research (published after 2016)[6] in the field of text analytics (with main object of analysis - online chat communication) focuses on the following questions of interest:

- What has been accomplished by researchers in the field to better understand this type of data and what knowledge and benefits it brings for the business? What are the practical applications of the analysis of such data?
- What are the specific characteristics of this type of data which should be taken into account during the application of quantitative analysis techniques?
- Special attention is paid to the main techniques used for text data processing, as well as overall research method in the analysis of chat data.

Each research article is analyzed in detail and a **structured summary is created in a tabular format** according to selected key characteristics of the articles - for example, data source and language, applied pre-processing and modeling techniques, used language resources etc. Below are

[6] The literature review also includes one article published in 2015, due to the fact that it shares to a certain extent some of the goals set in the dissertation.

mentioned only **some of the most important conclusions** made as a result of the literature review of research focused on quantitative analysis and knowledge extraction from online chat communication.

The addressed problems and application areas in the reviewed research articles are very diverse including: process automation in contact centers; increasing the efficiency of dialogue systems; improving customer experience in various online communication and entertainment platforms; detection of online fraud and crime (for example, online identity theft); improving the education process by applying innovative techniques to facilitate learning etc.

Not surprisingly, more than half of the reviewed research papers analyze chats in English. Of course, the reason for this is the abundance of publicly available data in English, as well as the availability of many language resources for this language. **The literature review reveals that currently there are no studies devoted to the analysis of online chat communication in Bulgarian. The last outlines an opportunity for research progress in this direction.** The most common sources of chat communication used in research are various social networks and forums (chat rooms). Data generated in a contact center is analyzed in a small number of the reviewed articles.

Regarding the main analytical techniques used in current research - the review reveals that after 2018, word embeddings became more commonly used in chat data analytics. Among the classical machine learning approaches, the most frequently used are the Naïve Bayes model and Support Vector Machines (SVM). In some of the most recent research articles are applied Transformer models. In topic modeling, the LDA algorithm or its different variations are among the most widely applied on chat data.

More than half of all the reviewed articles rely on the use of a variety of language resources. Some studies strongly depend on the usage of such resources not only in the preprocessing stage but also in the subsequent statistical analysis of chat data. The last means that replicating experiments on data in other languages is difficult, if not impossible, as similar language resources may not exist for them. Moreover, it is not surprising that there are many "manual" tasks (requiring human efforts) performed in the experiments – such as spell checking, annotation etc. Future research in the field of text analytics should focus on discovering new methods to facilitate the whole analytical process and to overcome to some extent the difficult and cumbersome "manual"

tasks during data analysis. The application of transfer learning is a step forward, but only one of the articles included in the review utilizes such methods - [10].

The review reveals that few studies propose an implementation of the applied methods and techniques in the form of an analytical tool (system) that could be used for text data analysis automation. Only two of the reviewed articles propose such analytical tool - [12], [19]. **It can be concluded that more research has to be done with the aim of developing comprehensive analytical tools for chat data analysis and adoption of effective techniques for text data visualization.**

As a result of the literature review, it can be concluded that **there are only two studies that share some of the main objectives and problems addressed in the dissertation. The first study focuses on the application of sentiment analysis for measuring customer satisfaction with the services provided in a contact center -** [16]**. The second article studies the most frequently asked questions/topics in online chat communication, conducted in the contact center of a banking institution -** [8]**.** The aim in [8] is to extract key client phrases and requests, using a variety of language resources for the Russian language. However, in order to be able to apply the authors' approach to text data in Bulgarian (or other language), it is necessary to have available similar language resources as those used in [8]. The detailed analysis of recent research in the field of chat data analysis facilitates the development of the methodology and some of the hypotheses addressed in the current study.

2.1.2. Text analytics in Bulgarian

One of the main aims of the dissertation is to present an up-to-date picture of the development in the field of text analytics in Bulgaria. The extent to which the application of various techniques for text data analysis is possible and accessible varies for different languages. The availability of NLP tools and various language resources is one of the factors determining the existing division between the so called "low-resource languages" and "high-resource languages". Of the latter, not surprisingly in the first place is the English language for which there are countless corpora, NLP tools, dictionaries, specialized software systems for text analysis, etc. The Bulgarian language is characterized as a "low-resource language", as a small number of linguistic resources are available for the purposes of its processing and analysis. After the review of current research focused on data analysis and information extraction from chat communication, in the dissertation

is conducted a **review, analysis and synthesis of key research in the field of text analytics in Bulgarian in two main areas.**

First, research in the practical area of natural language processing is in focus – outlined are the **main language resources which are available for Bulgarian**. The scope of this task includes various instruments or whole systems developed to ease textual data processing and analysis. Examples of such NLP tools commonly used in text analytics projects are different types of parsers, POS tagging and Named Entity Recognition (NER) tools, stemmers etc. When reviewing research focused on the development of NLP tools for Bulgarian of main interest is the applicability of such instruments and whether they are accessible by a web interface, software program or implemented in established software for statistical programming as R or Python. Special attention is being paid to the development and availability of text corpora in Bulgarian (and their type, size, text data domain and level of annotation).

The second central point in the literature review, analysis and synthesis is the **practical application of text analytics in Bulgarian to solve various economic or business problems.** Outlined are key research articles focused on the statistical analysis and experimentation with text data in Bulgarian for solving real-world problems. Under review are studies in three main practice areas of text analytics which, of course, have many intersections – document clustering, document classification, and information extraction. The review is mainly focused on: the business/economic problems being addressed; text data domain; applied analytical methods; availability of language resources as datasets, models or programming code for experiments provided as a result from the study.

Drawn are valuable conclusions regarding the degree of development of the field, the availability and applicability of language resources for the Bulgarian language and the extent to which text analysis has been applied on Bulgarian text data in practical business and economic problems. The scope is not limited to reviewing only the most recent research (for example, papers published in the last five years) - instead, the focus is on key research in the field. The literature review is comprehensive without being, or claiming to be, exhaustive. To the best of the author's knowledge, **this is the first attempt to outline key research in the field of text analytics for Bulgarian** in the two directions described above. Below are mentioned only **part of the most important conclusions** made as a result of the extensive literature review.

The global development of text analytics and more specifically – NLP, goes through three main stages. Initially, mostly rule-based approaches were used. Such methods despite being easier to understand are time-consuming and prone to human errors and do not generalize well. After this period, **the application of statistical approaches to NLP tasks began**. Machine learning methods are **far more reliable than rule-based approaches** since statistical inference is used to interpret and detect patterns in textual data. However, there is one important prerequisite – the availability of data. Such algorithms "learn" by utilizing annotated training data. **The third stage of NLP development occurred in the recent years with focus on transfer learning and utilization of Transformer models** [20]**.**

Despite being somewhat shifted in time, the development of NLP and text analytics for the Bulgarian language quite naturally follows the same path. The application of statistical approaches in experiments starts relatively later if compared to NLP development in general. The latter could be explained by the lack of annotated data at that point of time. The present review reveals that most commonly applied are methodologies based on machine learning approaches different from deep learning. So, one **direction for future research is the utilization of deep learning techniques** in both NLP tools development and applications with more practical focus. However, it should be noted that the last depends to a very large extent on the amount of available data - traditional neural network-based methods require a lot of data in order to achieve high efficiency.

In this regard, **the advent of transfer learning may be a workaround solution for such low-resource languages as Bulgarian**. Transfer learning allows the utilization of models trained on huge amounts of data which can be fine-tuned for specific tasks and languages. Such fine-tuning can be performed with significantly smaller amount of data. Without doubt **another direction for future development is the study of transfer learning and how it can be utilized for solving text analytics problems in Bulgarian**. Studies already focused on this hot topic are - [21], [22], [23], [24].

The review reveals that so far, **existing text corpuses in Bulgarian are few**. Among them the main are – two monolingual corpora annotated on different linguistic levels (BulTreeBank corpus and Bulgarian National Corpus - BulNC), two wordnets (BulNet, BTBWN) and a few bilingual/multilingual parallel corpora. The BulTreeBank corpus is characterized by high quality annotation at various linguistic levels and many key research articles use it in the development of various NLP tools. **The existing language resources (systems and tools) for Bulgarian** cover a

wide range of **natural language processing tasks**, including: **tokenization, stemming and lemmatization, POS tagging, dependency parsing, NER, word sense disambiguation, constituency parsing etc.** However, not all of these resources are available and easy to use and implement in real industry applications. The work of Popov, Osenova and Simov [25] marks the beginning of research efforts put into the integration of NLP tools for Bulgarian with open-source software like Python.

Currently, language resources for Bulgarian, if implemented at all, are "living" separately on different platforms/specialized software programs. **What can be said for sure is that future developments in the field depend on and should include integration with software such as Python, R or NLP frameworks like spaCy, NLTK, TextBlob, etc.** Integration of language resources with such software would help to assess their applicability in the industry. Naturally, if such integrations exist, this will ease practitioners and lead to more experiments and research with a practical focus.

The review reveals that **main research efforts in the text analytics field for Bulgarian start in the beginning of the 21st century**. There are not many research articles focused on the applications of text analytics on practical economic/business problems - more than half of the reviewed articles are published in the last 4 years (from 2016 to 2020). It should be noted that there is a lot of research and progress in the field of biomedical NLP for the Bulgarian language. There is a clear scientific interest and progress in the detection of toxic and misleading behavior in community forums and fake news detection.

The literature review reveals that almost half the studies focused on practical business/economic problems analyze news data. More should be done in the direction of human behavior analysis and social networks communication analysis (for example, using data from platforms such as Twitter, Facebook, BG-Mamma etc.). More efforts should be put in the direction of sentiment analysis which is extremely valuable in the analysis of political and social events [26], [27] and has many applications in the financial [28], [29], [30], accounting [31] and economic fields [32].

Various social, political or business problems are addressed with the help of text analytics in Bulgarian: website optimization; detection of "manipulation trolls" misleading public opinion on the Internet; sentiment analysis applied on user feedback; fake news detection; customer behavior analysis and others. The future of text analytics in Bulgarian should gradually shift to application

of more advanced and resource-rich NLP tasks such as question-answering, summarization, reading comprehension, relation extraction, semantic textual similarity, conversational AI.

An important finding made as a result of the extensive literature review, analysis and synthesis is that currently there is only one research article devoted on sentiment analysis in Bulgarian [33] **and there are no studies focused on the application of topic modeling techniques.** These two analytical tasks are central to the current study. In addition, **to the best of the author's knowledge, currently there are no published articles devoted on the analysis and information extraction from online chat communication between clients and employees (in Bulgarian).** The current study is the first for this combination of text data language (Bulgarian), data domain (online chat communication in the customer service domain) and analytical task (sentiment analysis/topic modeling). Exploring the available best approaches that could be applied, as well as potential issues, **the dissertation contributes to the body of literature devoted on the practical applications of text analytics in Bulgarian.**

The literature review reveals the need for development of modules for automation and facilitation of chat data processing and analysis. In this regard, **one of the contributions of the dissertation is that it provides the opportunity specific procedures of the established overall methodology for data processing and analysis (in Chapter II) to be replicated by other researchers in the field and used to automate and facilitate the analysis of online chat communication in Bulgarian.**

2.1.3. Topic modeling and sentiment analysis applied on text data

The dissertation also includes a literature review of the main approaches, current trends and established techniques for topic modeling and sentiment analysis of text data. **Topic modelling** (also known as topic discovery or topic extraction) is a popular unsupervised technique in the field of text analytics which is most often used to explore the structure of unknown corpuses of text documents. The concept of topic discovery has been known since 1990 [34] and has applications in numerous fields [35] - analysis of historical documents, scientific publications, fiction, source code analysis, event detection, image classification, opinion mining etc. In [36] is provided the following definition: *"Topic modelling is the general term behind a large group of algorithms used to reveal, discover and annotate thematic structure in collection of documents".*

The following definition explains the meaning of "topics" in the context of this type of analysis [34]: *"In topic modeling, the word "topic" specifically refers to a probability distribution over words part of a given text corpus, while also being used in the context of its more "general" meaning - a subject of discussion."* Put into simple words, topic modeling is the process of grouping words into topics based on their joined appearance in the text corpus being analyzed. That is, topics are collections of words that could be connected in a meaningful way [34].

In the context of one of the subjects of the dissertation, topic modeling provides an opportunity to draw interesting conclusions and identify trends in customer behavior by analyzing the main topics of interest to clients, as well as existing problems with goods and services. Among the main techniques for topic modeling are LSA (Latent semantic analysis), NMF (Non-negative matrix factorization), pLSA (probabilistic Latent Semantic Analysis), LDA (most commonly applied by researchers), many modifications of LDA (Correlated Topic Model - CTM, Pachinko Allocation Model - PAM, Author Topic Model - ATM) and others.

Regarding the other subject of the dissertation (customer satisfaction) - in the literature, the analysis of customer satisfaction with contact center services is most often carried out with the help of qualitative approaches [37], [38]. On the other hand, quantitative methods allow automation, faster analysis, and have more comprehensive coverage. Such methods might be considered as an important complement to the qualitative approaches, which are characterized by greater detail and accuracy. The current study focuses on the application of quantitative methods for measuring customer satisfaction. **The task of predicting whether a customer will rate his experience with the provided contact center chat service as positive or negative could be formally defined as a sentiment analysis problem.**

Formal definition of "sentiment analysis" [39], [40] – *"Sentiment analysis is a scientific field that deals with the analysis of opinion, sentiment and subjectivism, expressed in text, by using the computing power of modern technologies. This is a large problem area focused on the analysis of feelings, evaluations, attitudes, and emotions of people expressed towards various topics – goods, services, events, topics of social importance, individuals etc."*

In the context of sentiment analysis, most often is tackled the task to determine the sentiment polarity of a given text. What exactly is meant by "polarity" of sentiments expressed in textual data, Pang and Lee define as follows [39] – *"The task of determining whether a text expressing an opinion contains a generally positive or negative sentiment is defined as a task of determining the*

polarity of the text." **The subject of the dissertation is client's satisfaction, expressed in its polarity.** By "polarity" is meant specifically whether customers have rated positively or negatively their online chat communication with the contact center. **The dissertation tests empirically the assumption that text data in the form of chats contains signals sufficient to predict the opinion (sentiment) of the client about the provided chat service.**

In the existing literature, most applications of sentiment analysis are mainly on product/service/movie reviews, as well as on data from social networks (e.g., Twitter). The current study applies sentiment analysis in an area that is much less represented in the literature - online chat communication, generated in a contact center (the customer service domain). Such data imposes many challenges related to data processing, level of analysis, extent to which it contains expressed opinions/sentiments, presence of "irrelevant" information/noise (e.g., greetings at the beginning of the conversation, customer identification process and similar features part of the customer service process), which might additionally complicate the analysis. **The dissertation contributes to the existing literature by drawing valuable conclusions regarding such data characteristics and their potential effects on the application of sentiment analysis.**

Three main analytical approaches could be utilized in the sentiment analysis task [41]. The first one is based on the usage of sentiment lexicons (SentiWordNet [42], AFINN [43], VADER [44], SocialSent [45] and others) - such approaches do not require training data. Sentiment lexicons contain lists of words and phrases that are often associated with positive or negative sentiment. To the best of the author's knowledge, there is only one sentiment lexicon for text data in Bulgarian which is freely available to the research community [33]. **However, it is important to note that this lexicon has been developed on text data in a significantly different domain from that of the data which is a subject of the dissertation.**

The second approach to sentiment analysis is based on methods in the field of machine learning, while the third approach is a hybrid - a combination of sentiment lexicons and machine learning. Among the most popular classical algorithms for text data classification are support vector machines (SVM)/support vector classifier (SVC), Naïve Bayes model and logistic regression [46]. Gradually, the use of Transformer models [20], combined with transfer learning [47], is introduced in the field.

2.2. Methodology (Chapter II)

The steps in the methodology development are documented in [48]**,** [49]**,** [50]**.** In regard to the main aim of the dissertation are formulated the following four sub-objectives:

1. The first sub-objective is to create a module that **processes the data and structures it** in a format suitable for further analysis (Module I). The goal is to bring the data into a form that facilitates the application of quantitative methods for analysis.
2. The second intermediate objective is to develop a module that performs **extraction and analysis of the main topics of client interest** discussed in the online chat communication with the contact center (Module II).
3. The third sub-objective is to create a module that **analyzes the satisfaction (attitude/sentiment) of the client** with the provided contact center chat services (Module III).
4. The fourth intermediate objective focuses on the **business applications of the developed system** and the summarization and visualization of the information obtained in Modules II and III, as well as other interesting and important indicators that could be derived from the studied type of data (Module IV).

Each of these sub-objectives is related to the development of a specific module of the automated analysis system, which is devoted on a particular analytical task (Figure 1):

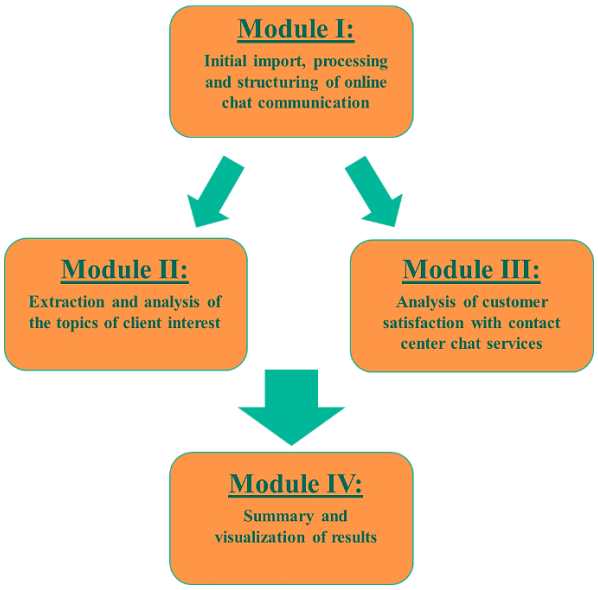


Figure 1. Scheme of the main modules in the system for analysis of online communication with customers

This system is developed **and could be implemented in a real environment with the help of the Python programming language**, which is used in all analytical experiments performed in the dissertation[7]. Among the main libraries utilized in the empirical study are: Gensim, scikit-learn, NumPy, pandas, Plotly, wordcloud, NLTK, treetaggerwrapper, BulStem.

2.2.1. Module I

In Module I is applied an algorithm for proper data import and structuring (Figure 2), since the raw data (Figure 3) does not allow the application of any analytical techniques. Part of Module I is also an algorithm for initial text data processing and normalization (Figure 4). This step is necessary due to some specific data characteristics and peculiarities, which if not taken into account, would reduce the quality of all subsequent analyzes. The output of this module is technically correct and structured data, which is used as input in all other modules of the system.

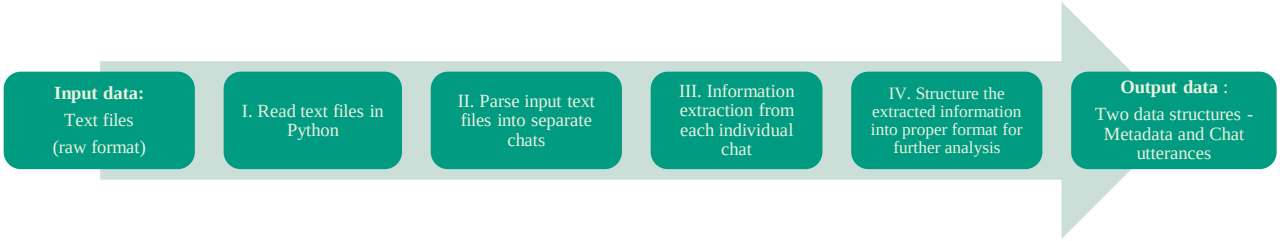
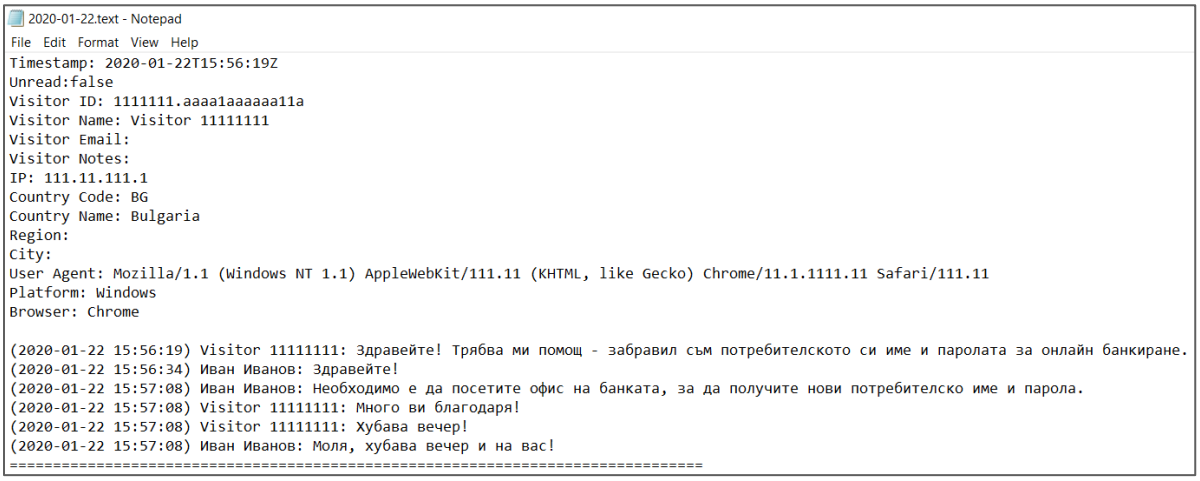


Figure 2. Diagram of the algorithm for proper import and structuring of chat data


2020-01-22.text - Notepad

File Edit Format View Help

```
Timestamp: 2020-01-22T15:56:19Z
Unread:false
Visitor ID: 1111111.aaaa1aaaaaa11a
Visitor Name: Visitor 11111111
Visitor Email:
Visitor Notes:
IP: 111.11.111.1
Country Code: BG
Country Name: Bulgaria
Region:
City:
User Agent: Mozilla/1.1 (Windows NT 1.1) AppleWebKit/111.11 (KHTML, like Gecko) Chrome/11.1.1111.11 Safari/111.11
Platform: Windows
Browser: Chrome

(2020-01-22 15:56:19) Visitor 11111111: Здравейте! Трябва ми помощ - забравил съм потребителското си име и паролата за онлайн банкиране.
(2020-01-22 15:56:34) Иван Иванов: Здравейте!
(2020-01-22 15:57:08) Иван Иванов: Необходимо е да посетите офис на банката, за да получите нови потребителско име и парола.
(2020-01-22 15:57:08) Visitor 11111111: Много ви благодаря!
(2020-01-22 15:57:08) Visitor 11111111: Хубава вечер!
(2020-01-22 15:57:08) Иван Иванов: Моля, хубава вечер и на вас!
================================================================================
```

Figure 3. Structure of an example chat in raw format.

[7] Two integrated development environments are utilized – Spyder and Jupyter Notebook.

Figure 4. Initial processing before using the chat data as input to Module II/Module III

The techniques applied in Module I are carefully chosen with respect to the specific characteristics of chat data used in the empirical experiments in Chapter III (in which the developed system for chat data analysis is tested empirically). The data is generated in the contact center of a large financial institution in Bulgaria. In order to use the developed system for the analysis of chats generated in a different chat platform (and respectively in different raw format), it might be necessary to modify some of the techniques in Module I to ensure proper data import and structuring.

2.2.2. Module II

In Module II, a combination of analytical techniques is applied with one ultimate goal - extraction and analysis of customers' main topics of interest discussed in the communication with the contact center. Chats are significantly different than data generated in other communication channels, not only in terms of their structure and linguistic features, but also as a source of information. One challenge in the topic modeling process is the presence of greetings, template messages used by operators in the process of client identification and service etc. These parts of the dialogue do not add any additional knowledge, but rather represent an irrelevant to the analysis information.

In this module are applied additional text processing techniques chosen with regard to the specific analytical task (Figure 5). **The proposed comprehensive research method in Module II is an author's combination of various methods and techniques carefully selected in order to achieve the research objectives**. It is important to note that some of the analytical procedures in topic modeling are improved, compared to those utilized in the initial experiments described in two of the publications related to the dissertation - [48], [49].

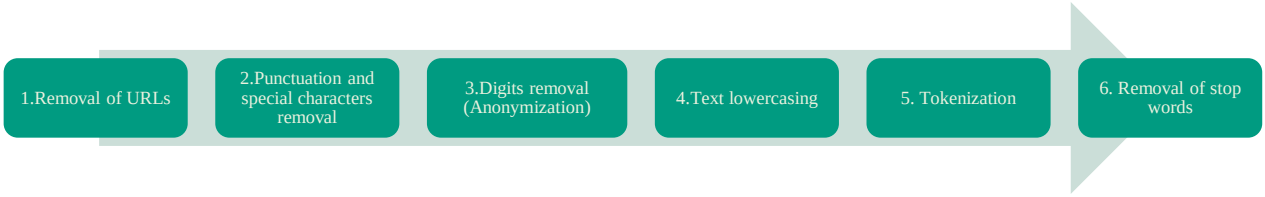


Figure 5. Text processing pipeline in Module II

The LDA algorithm is used for modeling and analysis of clients' main topics of interest. This choice for an algorithm is not surprising and is based on the huge volume of up-to-date research articles that apply it in a variety of experiments. To the best of the author's knowledge, according to the literature review in Chapter I, **the performance of this algorithm when applied on similar to the object of the current study type of data has not been studied before** (customer service via chat in the financial domain). **The dissertation aims to fill this gap in the literature and to draw conclusions about the performance of LDA on such type of data, the limitations and advantages of the approach, as well as data specifics that would affect the topic modeling results. In this regard is formed Hypothesis 1 (see section 1.4).**

Chat communication in a contact center represents a mixture of client and operator utterances. Each chat is usually initiated by the client who asks a particular question or presents his problem, asking for help. Following are some diagnostic operator's questions, completion of the client request and end of the conversation. **In this regard, arises the following question: "Do operators' utterances help in the task of extracting the main topics of client interest or rather hinder it by adding extra noise (irrelevant information) to data?"** Taking into account the opinion of experts in the contact center of the financial institution, on whose data the automated analysis system is applied in practice, Hypothesis 2 is formed (see section 1.4). To test this hypothesis, three levels of chat data representation are created:

1. Sample of **full-text** chats ("Sample I").
2. Sample of **all client utterances (only)** in a chat ("Sample II").
3. Sample of **beginning (first) client utterances (only)** in a chat ("Sample III"). The aim is to include only the customers' utterances in which they explicitly describe their problems and requests.

Although there are scientific papers expressing similar ideas in other research areas (e.g., comparison of the results and effectiveness of topic modeling applied to full and abstract text

[51]**), to the best of the author's knowledge, there are no studies analyzing the impact of chat data level of representation on the behavior of topic models such as LDA or similar.**

By applying LDA, each document in a corpus is represented as a random mixture of hidden (latent) topics, and each topic is represented as a different probability distribution over all words in the corpus [52]. In the dissertation, the LDA algorithm is applied to the set of individual words (known as "unigrams") contained in the text **on each level of chat data representation**. The utilized set of words (also called a "vocabulary") is directly related to the results and quality of the extracted topics. In this regard, **various filters are applied on the vocabulary** used in topic modeling in order to achieve optimal results. An example of such filter is changing the values of a parameter that controls the inclusion of words that occur very frequently or rarely in chats. The experiments include filtering the vocabulary based on an additionally developed stop words list specific to the case and data under study, as well as filtering words based on part of speech tagging.

Although topic modeling in the current study is performed in an unsupervised way, the research method includes relevant metrics for results evaluation, hypotheses testing and selection of the optimal number of topics. These metrics are **perplexity** [52] and $\boldsymbol{C_v}$ **coherence** [53]. The advantages and limitations of each of these two metrics are analyzed in detail. In the results evaluation are considered both the **author's interpretation** and the **expert opinion of employees** in the contact center of the financial institution, on whose data the automated analysis system is applied in practice.

2.2.3. Module III

In Module III of the system is developed an **author's approach for analysis of customer satisfaction** with contact center chat services reflecting the challenges imposed by the object of the dissertation. In data used for the empirical test of the created system, a very small number of customers have provided a customer rating. This observation has been also made in other studies analyzing online chat communication in a contact center, and is considered as a specific characteristic of this type of data [16]. This outlines the need for an analytical tool for customer satisfaction analysis. **Of greatest interest to the business is to detect customer service dialogues from which the customer left disappointed**. However, there might be chats in which the client does not seem to have expressed an opinion/sentiment and in the end he/she still provided a bad service rating.

Based on findings in the detailed literature review and **taking into account the described specific data characteristics, in the dissertation is applied a machine learning approach for customer satisfaction prediction**. This way it is expected to build a model that captures the specifics in the written language used by customers in this domain – it is assumed that a statistical model which learns from historical data will detect some "invisible" signals in the communication that lead to bad service rating. This also implies achieving greater prediction accuracy than that which might be achieved by sentiment lexicons [46]. As mentioned earlier in section 1.4 the **empirical results and findings published in** [16] on the applicability of such lexicons to data in the customer service domain further justify this choice.

In the current study is tested empirically one very important assumption, namely that the **textual dialogue (chat) between a client and an operator, contains enough signals** that could be used to develop a customer satisfaction prediction model. Whether this assumption is realistic or not is one of the research questions in the dissertation, on the basis of which is formulated **Hypothesis 3** (see section 1.4). If text dialogues indeed contain such signals, then training a statistical model on them will lead to optimal results compared to those achieved by applying a "naïve forecast" (for example, a random guess model). **The usage of mathematical algorithms in order to "learn" from historical data should lead to better results only if the data really contains valuable signals regarding customer satisfaction.**

Additionally, in the task of customer satisfaction prediction, arises the following question: **"Is it necessary to use the whole dialogue or just part of it will be sufficient to predict the final service rating?"**. The analysis of findings made in [16] leads to the formulation of Hypothesis 4 (see section 1.4). **Results in this study indicate that sentiments detected in the last part of each chat (in both the customers' and operators' utterances) are the most important features in customer satisfaction prediction**. In this regard, the current analysis is applied on **three levels of chat data representation**:

1. Sample of **full-text chats** ("Sample 1").
2. Sample of **all client utterances** in a chat ("Sample 2").
3. Sample of **final (last) client utterances (only)** in a chat ("Sample 3"). In Sample 3 it is assumed that in the end of the conversation the client either expresses his gratitude or the opposite if the service is not satisfactory and his/her request is not fulfilled.

Following is a brief description of some of the main steps in the research method applied in Module III. First, based on the available data, **the target variable is formed** - chats in which the final rating of the client **is known,** are used as training data (about **13.6%** of all the available data). The problem addressed in Module III could be formally defined as follows:

Given a chat *C*, the task is to classify it into one of the following categories of service rating: *Good* (0) or *Bad* (1).

Text processing and vectorization are performed, taking into account both the data characteristics and problem being solved in Module III (Figure 6). An analytical procedure is developed in order to study the effect of the following parameters/techniques, with the ultimate goal to achieve a workable solution and optimal model performance in customer satisfaction prediction:

1. **Level of chat data representation** ("Sample 1"/ "Sample 2"/ "Sample 3").
2. **Text processing techniques** - testing different forms of stemming [54] applied on data and analyzing their effect on the results ("Stemming 1" / "Stemming 2" / "Stemming 3" / "original form of text data"). Details on the differences between the three forms of stemming are available in the main text of the dissertation.
3. Experimenting with **different sets of explanatory variables** extracted directly from text data. Chats, by themselves, form a set of potential **"textual" explanatory variables**. In addition to the words in chats, certain grammatical features of the text extracted with the help of a POS tagging tool are utilized in the prediction - **these are called "morphological" explanatory variables**. **The latter are used either in combination with the "textual" explanatory variables or as a method for their selection** (selected are words which fall in the category of specific parts of the speech). **The selection includes the following parts of speech - nouns, verbs, adjectives, interjections and adverbs**. POS tagging is applied using the TreeTagger tool [55].
4. **Machine learning algorithm** – the empirical study includes the application of three machine learning algorithms - support vector classifier (SVC), Bernoulli Naïve Bayes model (BNB) and logistic regression (choice based on a detailed review of studies analyzing online chat communication).

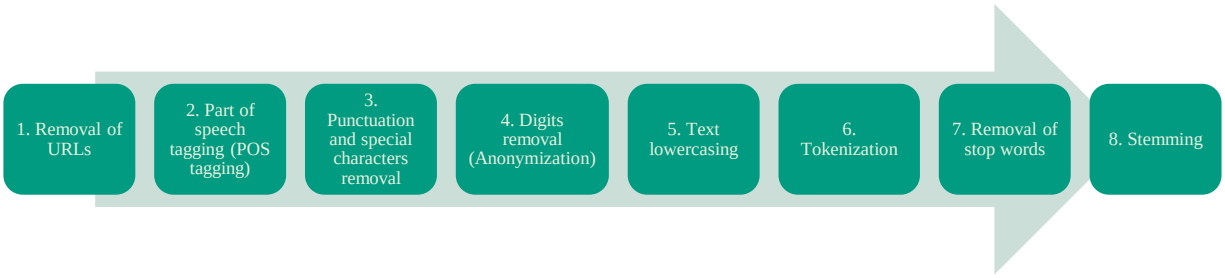


Figure 6. Text processing pipeline in Module III

The utilized method for model validation is ***k*-fold cross validation** [56]. In hypotheses testing and evaluation of the predictive power of the created models, several established metrics are used - **accuracy, precision, recall, F1, F-beta**. Based on the various combinations of tested parameters and with the help of the evaluation metrics, the model with optimal performance is found.

The application of deep learning methods, transfer learning and Transformer models, which are significantly different from the utilized classical approach, **falls outside the scope** of the empirical experiments. **The aim is to analyze what can be achieved in the task of sentiment analysis of online chat communication through the application of traditional approaches in the field of machine learning** and to make a comparison with the results achieved by other researchers who rely on the usage of sentiment lexicons for solving similar tasks [16].

A significant number of approaches based on artificial neural networks require a larger amount of data in order to achieve efficient performance. In the business case analyzed in Chapter III of the dissertation, currently this prerequisite is not fulfilled and the accumulated volume of data is not significant. On the other hand, transfer learning addresses precisely such scenarios, aiming to use the knowledge gained from similar data and tasks in other domains. Currently there is little research on the application of deep learning and transfer learning in the analysis of text data in Bulgarian - the field is yet to develop. This outlines one of the **future directions** of the current study – application of transfer learning and deep learning methods for text data representation as well as predictive modeling of customer satisfaction.

2.2.4. Module IV

The purpose of Module IV is to put an emphasis on the interpretation of the results from the applied analytical techniques in the other modules. The main aim is to shed light on the applicability of results and their usefulness and meaning for the business. **Module IV facilitates interpretation of the results from the analysis of online chat communication in a contact**

center, as well as knowledge extraction from such type of data, focusing on the business applications of the created system.

Module IV aims to illustrate how the results could be optimally utilized and represented in order to reach more effectively people in the business who are most interested in them. Module IV combines simple analytical and visualization techniques (for example, word clouds or graphs that track changes in a given characteristic over time). The extracted knowledge from data in Module II and Module III is combined, thus achieving a synergistic effect in the analysis and interpretation of results. The latter allows drawing valuable insights regarding the level of customer satisfaction and its connection with topics discussed during the communication with the contact center.

With regard to the specific banking institution, whose data is used to test empirically the developed system, at least several departments would be interested in utilizing the system - for example, Contact Center Department, Service Quality Department, Customer Relationship Management Department. The set of analytical techniques and visualizations in Module IV illustrates how the system will appear for its end (business) users. While Module I, Module II and Module III could be referred to as the "back-end" of the system, Module IV is its "front-end" – end users interact directly with this module and its purpose is to provide them the opportunity to find answers to the following questions:

- What knowledge is extracted from data?
- How this knowledge is valuable for the business?
- What business decisions could be made based on the knowledge extracted from data?

2.3. Empirical study (Chapter III)

In Chapter III, the automated analysis system is **tested empirically on real data** (in the context of a specific business case). In the empirical study is analyzed online chat communication between clients and operators in the contact center of a large financial institution in Bulgaria. A large part of the main results and conclusions made in the empirical experiments are documented in [48], [49], [50]. **As mentioned earlier, some of the analytical techniques have been improved compared to those used in the initial experiments.** The sample of data used for testing the system consists of **37 529 chats** conducted in the period from **22.01.2019** to **01.04.2021**. As different levels of chat data representation are used in the analysis, for convenience, the full sample of data is denoted as "Sample I".

2.3.1. Topic modeling (Module II)

Initially, the analysis of main topics of client interest is performed on Sample I (this sample contains all client and operator utterances). After that, a next experiment aims to test which level of chat data representation (see section 2.2.2) will lead to optimal results from topic modeling. The LDA algorithm is applied in an unsupervised setting - the actual number of topics in the sample is not known in advance. With the help of empirical methods is developed a procedure to determine the optimal number of topics – the C_v coherence metric and perplexity are used. The last aims to assess model performance on previously unseen data, while topic coherence aims to assess the quality of the topics in terms of their interpretability and meaning.

In this sense, topic coherence is closer to the "human" understanding and perception of topic model quality. C_v takes values between 0 and 1. Higher values of C_v imply that topics are more meaningful and easier to understand and, respectively - the model has better performance. Some of the fundamental studies in the field determine the following reference values for C_v - values below 0.4 indicate low to poor quality of the created topic model, while values between 0.6 and 0.8 are considered optimal [51], [57], [58]. Values above 0.9 are unlikely and considered rather suspicious[8].

A crucial step in topic modeling is choosing the vocabulary used in model development. In the empirical study are tested **five different forms of vocabulary filtering (on Sample I)**. Vocabulary filtering consists of various changes in the following parameters - **removal of the most common/least common words, making use of an additional manually created stop words list and feature (word) selection based on part-of-speech tagging**[9]. Table 1 compares the results from the application of LDA on Sample I (after the application of different vocabulary filters).

Based on an extensive analysis of the results obtained on Sample I, it can be concluded that models with 15 to 20 topics are characterized by optimal performance, both in terms of coherence of the obtained topics and in terms of model behavior on previously unseen data (according to the perplexity measure). It is important to note that the approach proposed in some fundamental studies in the field [51], [59] is applied in the current one in order to select the optimal

[8] The reference values concern the overall coherence of a topic model (i.e., the average of the sum of the individual coherences of each topic in the model).

[9] More details regarding the differences between applied vocabulary filters are available in the main text of the dissertation.

number of topics (according to C_v). A graphical analysis is performed in order to track model coherence for each number of topics (for each filtering of Sample I are analyzed models with 2 to 50 topics). The graphical analysis allows to easily distinguish the moment when coherence begins to vary around the same value.

Table 1. Sample I – comparison of the results after the application of LDA

Vocabulary filter type	*Number of words in the vocabulary*	*Chosen optimal number of topics*	*Average value of C_v for the chosen number of topics*	*Average value of C_v after the chosen number of topics*
Filter 1	13 396	15	0.6127	0.5756
Filter 2	13 177	20	0.5844	0.5617
Filter 3	8 556	15	0.6233	0.5716
Filter 4	8 497	20	0.5308	0.5247
Filter 5	**4 592**	**16**	**0.6655**	**0.6163**

The highest value of C_v is achieved after the application of Filter 5 on the vocabulary used in data modeling of Sample I - only nouns are selected using a tool for part-of-speech tagging in Bulgarian [55]. Both perplexity and C_v coherence indicate very good performance of a topic model with 16 topics - this is the final choice for the optimal number of topics presented in Sample I. The choice is also in line with business expectations regarding the number of general topics of client interest presented in the data.

After the application of the LDA algorithm on Sample I, each chat can be described by its topic distribution. **The words with highest probability to occur in a given "topic" can be used to determine what the topic is mainly about – these key words reveal the main ideas/concepts expressed in discussions part of this "topic".** The data analyst's job is to extract the meaning, logic between groups of key words and label/determine the "topic".

In the main text of the dissertation are analyzed in detail the extracted topics of client interest, their quality, keywords, level of association as well as chats in which they are dominant. A library for specialized interactive visualizations is used [60] in order to facilitate the interpretation of topics as well as the discovery of interesting patterns and associations between them. The library allows to assign the topics into more general groups ("clusters") and analyze their level of dominance in the sample of data. The analysis of the extracted topics leads to the distinction of four main groups of client discussions, namely:

1. **Lending and various steps in this process** - requirements for lending, lending process, bank decisions on loans/refinancing etc.
2. **Digital banking and online payments** - registration for online banking, signals for problems in online banking, receiving a message with code for online payments and others.
3. **Cash operations** – cash withdraw/deposit in a bank branch, questions regarding currency exchange (current exchange rates), searching for information about location/working hours of bank branches and others.
4. **Card products (credit and debit cards)** – account opening/closure (requirements and conditions), fees, payments and transfers, seizures of bank accounts etc.

None of the extracted topics is strongly dominant in client discussions, but the graphical analysis reveals that there is a slight dominance of requests and questions regarding the lending process and card products' fees. Following the extensive data analysis of Sample I, the next part of the empirical experiments aims to investigate which chat data representation will lead to optimal results in topic modeling (see Hypothesis 2, section 1.4). Section 2.2.2 describes the three levels of chat representation analyzed in the empirical study (Sample I, II and III). The samples are created by applying the same data processing techniques and analytical procedures.

In the topic modeling of Sample I are tested five different filters applied on the vocabulary. Although Filter 5 proves to be optimal for Sample I, for the sake of analysis completeness, the effects of some of the other filters (those that would be utilized more often in practice) applied on Sample II and Sample III are also studied during the empirical analysis. The values of the parameters in all the applied vocabulary filters are the same as those used in the analysis of Sample I. The results are analyzed graphically, and in Table 2 is provided a summarized comparative analysis between the results from topic modeling of Sample II and Sample III and the best topic model developed on Sample I.

Table 2. Sample II and Sample III – comparison of the results after the application of LDA

Sample	*Vocabulary filter type*	*Number of words in the vocabulary*	*Chosen optimal number of topics*	*Average value of C_v for the chosen number of topics*
Sample I	**Filter 5**	**4 295**	**16**	**0.6655**
Sample II	Filter 2	10 221	21	0.5268
	Filter 3	6 278	17	0.5363
	Filter 5	**3 725**	**15**	**0.5423**
Sample III	**Filter 2**	**5 191**	**15**	**0.4408**
	Filter 3	3 153	16	0.4393
	Filter 5	2 007	15	0.3987

Following is a summary and discussion of some of the most important conclusions based on the whole empirical analysis aimed at the discovery of topics of client interest:

- ✓ **The vocabulary used in topic modeling has a great influence on the results.** The different vocabulary filters might lead to considerable differences in the quality of the extracted topics, but **do not affect the choice of the optimal number of topics** - this number remains similar. **Filtering the vocabulary by selecting only nouns (Filter 5) using a parts-of-speech tagger** leads to **optimal results** in topic modeling applied on Sample I and Sample II. It can be concluded that nouns provide the most important context in relation to the topics discussed by customers.
- ✓ **The model with the highest quality of topics** is created on **Sample I** after the application of **Filter 5** on the vocabulary. Compared to the other filters, this one leads to the largest reduction in the vocabulary, which eventually proves to be optimal. These results demonstrate that **it is not the quantity of words in the vocabulary that is important, but their quality/relevance**. The selection based on part-of-speech tagging facilitates filtering out the noise in data and including only the most important words which in turn leads to a higher level of distinction between the extracted topics. Another benefit is the **optimization of the time required to train the model** while **lowering the chances of overfitting**.
- ✓ Nevertheless, **when choosing a specific vocabulary filter type it is important to take into account the size of the data sample (the total number of words in the vocabulary).** The application of Filter 5 leads to suboptimal results on Sample III, and the most probable explanation for this behavior is the drastic reduction in the number of words included in topic modeling. The reason for that lies in the fact that this sample is smallest in size compared to the other two samples since it contains only the first client utterances. In this regard, there are specialized methods aimed at **topic modeling of short texts** (such as tweets posted in the social network Twitter) [61]. In this case, such methods would be more suitable for application on Sample III, but continuing the experiments in this direction is beyond the scope of the dissertation and is considered as **one future perspective**.
- ✓ Based on the results obtained on Sample I, **Hypothesis 1 (see section 1.4) is confirmed**. Despite the specific characteristics inherent to chat data, the **traditional topic modeling algorithm LDA leads to the development of a topic model characterized by high coherence**. The coherence of the model trained on Sample I (Filter 5), as well as the extensive

analysis of the topics included in it, confirm that the LDA algorithm performs at an optimal level in chat data modeling and leads to achieving a **workable solution**.

- ✓ The comparative analysis between the three levels of chat data representation reveals that **optimal model performance is achieved by using Sample I** (regardless of the applied vocabulary filter type). Topic model coherence achieved by using Sample II or Sample III is significantly lower than that achieved when Sample I is used. This observation indicates a suboptimal overall model performance and lower level of interpretation and quality of the extracted topics (for Sample II and Sample III). The results on these two samples are also characterized by **greater variability**, which indicates a **lower degree of confidence and reliability of the created topic models**.
- ✓ Based on these results, **Hypothesis 2 is rejected** - the utilization of only client utterances **does not lead** to the development of a model with better performance compared to that achieved when operator utterances are included in topic modeling. Hence, it could be stated that **operator utterances provide an important context** that helps to extract topics of client interest and develop a workable solution.
- ✓ It is important to note that although the quality of the extracted topics decreases when topic modeling is applied on Sample II or Sample III, **the optimal number of topics remains the same** with some minor differences (again, the optimal number of topics is around 15-20). This result is in line with the expectation that regardless of the utilized chat data level of representation, the number of extracted topics will be approximately the same and they will be characterized by many similarities.
- ✓ One **future perspective** is to study the application of techniques such as stemming and lemmatization and their effect on topic modeling of text data in Bulgarian. Another **future perspective** is to study the opportunities provided by deep learning and transfer learning for topic modeling in Bulgarian, and the applicability of such techniques on the object of the dissertation.

In conclusion, the empirical study achieves all the research objectives in terms of topic modeling of online chat communication. Hypothesis 1 and Hypothesis 2 are empirically tested (see section 1.4) and an answer to one of the research questions in the dissertation is formulated. **The empirical analysis of the combination between the traditional topic modeling algorithm LDA, different levels of chat data representation, as well as different vocabulary filter types leads to the**

development of a comprehensive and efficient approach to topic modeling of online chat communication.

Drawn are valuable conclusions regarding different levels of representing online chat communication generated in a contact center – to the best of the author's knowledge, this topic has not been studied by other researchers in the field. The dissertation examines the impact of these levels of chat data representation on the results and especially on the quality of the topics extracted by the application of LDA. In addition, a **reliable approach for processing and normalization** of online chat communication is proposed (before topic modeling).

To the best of the author's knowledge, the **potential, performance, and limitations of the LDA algorithm when applied on online chat communication in a contact center (banking domain) have not been studied up until now.** Despite the obvious practical benefits of current research findings for companies in the banking domain, it is important to emphasize that the **developed system is applicable in any other industry** in which such type of data is generated as a result of the business processes.

It can be concluded that the current study proposes a **comprehensive research method** for topic extraction which overcomes some of the difficulties arising from the structural characteristics inherent to chat data. Unlike other studies with similar objectives, the current one proposes **a purely statistical approach** for topic modeling of chat data. The only linguistic resource necessary to apply the research method to similar data in another language is a part-of-speech tagger. **As a result, the proposed approach could be more easily replicated by other researchers in the field.**

2.3.2. Customer sentiment analysis (Module III)

Module III of the automated analysis system aims to establish a reliable research method for predicting customer satisfaction with contact center chat services. From Table 3 it becomes clear that only **13.66%** of all chats in Sample I (this sample represents all available data) are characterized by a client rating - the quality of the provided service in the remaining **86.34% of chats** is unclear. **The aim of Module III is to use historical information and machine learning to analyze the dependencies in communication between the customer and the operator, which signal whether the customer is satisfied or rather dissatisfied with the service provided.** For the purposes of developing a satisfaction prediction model, in the study is utilized a sample of chats

in which the customer's rating is known - this sample consists of **5 125 chats** conducted in the period from **28.01.2019** to **01.04.2021**. This sample containing the chats in their full format (i.e., including all utterances) is denoted as "Sample 1".

Table 3. Chat distribution in Sample I according to the availability of client rating

	Number of observations	*As %*
Chats with rating	5 125	13,66%
Chats without rating	32 404	86,34%
Total	37 529	100%

Customers can evaluate the service as either **"good" or "bad".** The target distribution is very unbalanced – from Table 4 becomes clear that the dominant category of feedback is "good" (this observation is rather expected). Interactions evaluated from customers as "bad" represent only **10,4%** of the sample. The unbalanced target distribution will most probably impact classification results. However, in the current experiment the original target distribution is retained, and no sample balancing techniques are applied to data.

Table 4. Target distribution - service rating

Target category (rating)	*Number of observations*	*As %*
Bad	533	10.4%
Good	4 592	89.6%
Total	5 125	100%

In the empirical experiments are tested Hypothesis 3 and Hypothesis 4 (see section 1.4). In connection with Hypothesis 3, the concept of "naïve forecast" is introduced - **this is a prediction that is not based on machine learning**. The "naïve forecast" approach is based on the idea of comparing the performance of a model that knows nothing about the data (does not extract knowledge from data and does not analyze the dependencies between observations) and a model that learns from the historical information provided by the data, in order to predict the value of the target variable.

Such a comparison will directly provide an answer to the question whether using machine learning to solve the analyzed problem is meaningful and useful. **As the best benchmark against which to compare the performance of the customer satisfaction prediction model is chosen a "naïve prediction" which takes into account the class distribution in data (this knowledge is known in advance).** By using this type of "naïve forecast", 10.4% of the chats are randomly classified as "bad" and 89.6% as "good" (this naïve prediction strategy is known as "*weighted*

guess classifier"). The values of the metrics in Table 5 are used as a benchmark in assessing the extent to which the application of machine learning on the object of the dissertation is useful in the task of predicting customer satisfaction.

Table 5. Values of the model evaluation metrics in the case of a "naïve prediction" (in the current use case)

	Accuracy	*Recall*	*Precision*	*F1*	*F-beta*
Naïve Forecast	~0.81	~0.104	~0.104	~0.1	~0.1

The aim of the analysis is to test various parameters/techniques of interest so as to obtain an optimal performance of the prediction model and to achieve the most efficient solution. The effect of the following parameters/techniques is studied during the experimentation with data and search for the optimal solution (more details are provided in section 2.2.3) - 1. **Level of chat data representation** (related to Hypothesis 4 - see section 1.4) - Sample 1, Sample 2 and Sample 3 are created; 2. **Text processing techniques** (stemming); 3. Experimenting with **different sets of explanatory variables** ("textual" and "morphological"); 4. **Machine learning algorithm**.

Developed is an analytical procedure which integrates all possible combinations among the selected four parameters. It is important to note that the same text processing techniques are applied to each level of chat data representation (Sample 1, 2 or 3) before the data modeling stage. A total of **216** prediction models are created (**3** levels of chat data representation X **4** different forms of stemming applied to text data X **6** different sets of explanatory variables X **3** different prediction algorithms). It is important to note that the research method applied in the empirical study is improved compared to the initial experiments published in [50] - 1. A better suited model validation procedure is chosen; 2. Each form of stemming applied to data (including the text in its original form) is included as a separate parameter during modeling; 3. The experiment is extended - two more machine learning algorithms are included (Bernoulli Naïve Bayes and SVC).

A total of 72 models are created for each level of chat data representation. **Modeled is the probability that a given online chat service will be rated as "bad".** In the evaluation of model performance, more weight is put on the recall and F-beta metrics, as from a business viewpoint it is more important to train a classifier that captures accurately dialogues which provoked **client dissatisfaction**. More emphasis is placed on the model's ability to correctly identify all observations that fall into the "bad" rating class, even if this might be to some extent at the expense

of model's precision. To facilitate the comparative analysis, Table 6 presents the three models with optimal performance (according to F-beta) for each level of chat data representation, respectively.

Table 6. Models with optimal performance according to F-beta (for each level of chat data representation)

Sample	*Algorithm*	*Text processing technique*	*Set of explanatory variables*[10]	*F1*	*Accuracy*	*Recall*	*Precision*	*F-beta*
Naïve forecast				~ 0.1	~ 0.81	~ 0.104	~ 0.104	~ 0.1
Sample 1	SVC	Original form	N-gram model + POS tags	0.8921	0.8972	0.4089	0.5059	0.4250
Sample 2	SVC	Stemming 1	N-gram model + POS tags	0.884	0.889	0.383	0.457	0.395
Sample 3	**BNB**	**Stemming 3**	**N-gram model**	**0.8630**	**0.8533**	**0.4784**	**0.3494**	**0.4450**

Following is a brief summary and discussion of some of the more important conclusions made during the empirical analysis dedicated to customer satisfaction prediction:

- ✓ **Hypothesis 3 is confirmed** - the application of machine learning to predict customer satisfaction leads to optimal results compared to those obtained by using "naïve" approaches that are not based on techniques for knowledge extraction from data. Each of the models characterized by optimal performance (for a given level of chat data representation) performs better in the prediction task, compared to what can be achieved with a "naïve" approach.
- ✓ The applied empirical techniques allow the formulation of an answer to the second research question, which has a central role in the dissertation (see section 1.4). **Chats contain enough signals regarding customer satisfaction and can be utilized in the development of a predictive model which might serve as a reliable basis for developing a working business solution.**
- ✓ The finally selected model, characterized by **optimal performance** compared to all the others (according to F-beta), is created only with the help of the customers' final utterances (Sample 3), on which "Stemming 3" is applied. This model is developed using the Naïve Bayes algorithm, applied to a set of unigrams and bigrams extracted from the text. **Hypothesis 4 is confirmed** – utilizing a sample consisting of clients' final utterances in the communication leads to optimal prediction results (according to the F-beta metric, which has the greatest weight in the evaluation process).

[10] The N-gram model utilizes the so-called unigrams (single words) and bigrams (two consecutive words) extracted from the text. Information about the presented parts of speech in the text can be also added to this model. More details on the utilized sets of explanatory variables are available in the main text of the dissertation.

- ✓ It can be concluded that the **final customer utterances contain the most valuable information and important signals in terms of customer satisfaction**. This conclusion is supported by the fact that according to some of the used evaluation metrics, the performance of models trained on Sample 3 is at the same level, if not better than that achieved by utilizing other levels of chat data representation. One advantage of using Sample 3 is that it is characterized by a larger feature selection, as it contains only the final customer utterances. This reduces the likelihood of model overfitting, as well as shortens the model training time. It can be concluded that much of the redundant information in data (which does not add valuable knowledge) is excluded at this level of chat data representation.
- ✓ **The confirmation of Hypothesis 4 and the conclusions made in the current study are in line with the results of other researchers in the field.** Park et. al. [16] analyze which are the explanatory variables with the highest predictive power in customer satisfaction prediction. Their results indicate that the customer's sentiment expressed in the final chat utterances is among the most important variables in the prediction task.

To conclude, one of the main goals of the current study is accomplished by applying a **purely statistical approach** for sentiment analysis in the customer service domain. A research method for customer service rating prediction based solely on chat data is developed and tested during the empirical study. The problem is studied from different angles by developing three distinct chat representations and conclusions regarding their usefulness in the prediction task are formed. **To the best of the author's knowledge, the current study is the first explicitly aimed at service rating prediction based solely on textual features combined with grammatical information extracted from chat data.** The study also contributes to the research field devoted on the practical applications of text analytics in Bulgarian aimed at solving economic and business problems.

It is important to note that the proposed approach does not rely on sentiment dictionaries which makes it suitable for low-resource languages as Bulgarian. The current study is the first for this **combination of task, text data language, and domain**. The study is among the few devoted on sentiment analysis applied on customer service dialogues and contributes to the research in the field of sentiment analysis for low-resource languages. The use case under study is in the banking domain but the proposed **research method, analytical techniques and approach for interpretation of results and findings** might be helpful in any other domain in which similar

communication is being analyzed. In addition, the results and conclusions might be useful to other researchers in the field who analyze similar type of data.

One of the **future research directions** concerns the application of transfer learning and Transformer models. A future perspective is to investigate the availability of pretrained models suitable for text data in Bulgarian and if such are found - to assess the extent to which these models are applicable to the object of the dissertation and the problem at hand. An interesting direction for future development is the analysis of whether such approaches could be combined with the techniques illustrated in the dissertation, so as to obtain an even more accurate prediction of customer satisfaction. Also of interest are the opportunities for applying deep learning methods and their comparison with the classical machine learning approach utilized in the dissertation. Of course, satisfactory results from the application of methods based on artificial neural networks would be expected in the event that a larger amount of available data is accumulated in the presented use case.

Another direction for future development is based on some of the empirical results published in [16]. The authors conclude that metadata associated with chats between customers and operators (e.g., chat duration, operator response rate, time of the day, number of utterances, etc.) has low predictive power in the satisfaction prediction task. A future direction for development is to investigate the effect of using metadata as explanatory variables and analyze whether this will improve the performance of a prediction model based only on "textual" variables. For example, it would be valuable to examine whether the slower average response rate of the operator, as well as maximum delay are factors in leaving a "bad" rating. An interesting direction similar to the research ideas illustrated in [21] is to compare the prediction results obtained using only structured data (metadata) and those obtained using only unstructured data (chat text).

2.3.3. Summary and visualization of results (Module IV)

Module IV combines additional analytical techniques for knowledge extraction from data, the main purpose of which is to:

- ✓ Facilitate the interpretation and acquisition of more context on the extracted topics of client interest.
- ✓ Track the development of topics over time.
- ✓ Assess the level of complexity of the discussed topics/client requests.

- ✓ Combine the knowledge extracted in Module II and Module III, thus obtaining a synergistic effect of the analysis of customer satisfaction as well as topics of interest.

Among the applied analytical techniques in Module IV are: 1. Graphical analysis of topics with the help of "word clouds"; 2. Development of **indicators measuring the complexity of customer requests** falling into a particular topic (by utilizing metadata associated with each chat in the sample - for example, the chat duration, number of utterances, response rate of the operator etc.); 3. Analysis of customer satisfaction in discussions on various topics; 4. Extraction of key phrases from the discussions related to specific topics; 5. Graphical analysis of the development of topics over time (detection of peaks in particular types of requests and tracking trends among customers' interests).

Module IV illustrates only part of the possible applications of the system for online communication analysis. There are many directions for improving the system and adding a variety of functionalities helpful to the business users. Potentially, the system could be expanded to include data from other communication channels. In the first place, the managers in the Contact Center department of a given company would benefit the most from such system since it provides direct measures of operators' performance and the level of coverage of various requests. This knowledge could potentially lead to a number of optimizations in contact center processes. Apart from that, in the recent years there has been a trend to implement automated dialogue systems in order to optimize customer service. In this regard, topic modeling and key phrases extraction applied to dialogues might be useful in at least two directions.

First, the system provides an assessment of the complexity of customer requests addressing specific topics. This knowledge might be used to assess which types of requests have the greatest potential to be easily automated, and vice versa - which requests will be the most difficult to handle and would require supervision by operators. Second, key phrases extraction as well as similar analyzes (for example, extracting synonymous client phrases) might be very useful in the process of chatbot development (common customer phrases could be used as training data).

Another business unit in the company that might be interested in the developed system is the Service Quality Department. It would be valuable to analyze when the client is more likely to be disappointed and the reasons behind this (such discussions could be analyzed in detail). This knowledge might help in identifying "problematic" topics for clients and directions in which the services need to be improved. Companies often use surveys to analyze customer opinion. However,

this approach is based on what the company finds as important and interesting. By utilizing only this approach, some important topics might remain in the "blind spot" of the business. Exactly this outlines the importance of the analysis of communication between a customer and an operator in a free format – this may lead to the discovery of important topics and issues overlooked by the company.

The Customer Relationship Management Department is another business unit that might also benefit from the proposed system. A simple search based on key words could be used to identify customer problems with specific products, services and topics. The system could also help in tracking the success of marketing campaigns and various company initiatives. In addition, the knowledge gained from the communication with clients could help in customer profiling and preparation of special offers.

In conclusion, the analysis of customer satisfaction as well as topics of interest combined with the analysis of metadata available for each chat, lead to many synergies and extraction of additional knowledge from the data. Only part of the number of benefits for the business are illustrated with the help of different examples of departments that would benefit from the functionalities of the system for online communication analysis. Thus, the results achieved in the dissertation are presented not only from an analytical point of view, but also from the viewpoint of their significance and interpretation for the business. **An emphasis is put on the applicability and practical meaning of the results.**

III. Conclusion

In conclusion, the main aim of the dissertation is achieved by applying a research method based entirely on quantitative methods for textual data analysis. An automated system for analysis of the main topics of client interest as well as customer satisfaction with the services provided in a contact center in which communication is in Bulgarian is developed in the current study. The structure of the system is illustrated, the methods for its development are presented in detail, as well as its specific applications in the business. Furthermore, an up-to-date picture and detailed analysis of the research progress and future directions for development in the field of text analytics in Bulgarian are also provided. The empirical study provides an answer to the two research questions underlying the dissertation. Each of the hypotheses defined in relation to these two research

questions is tested using carefully chosen metrics to evaluate the results obtained and the effectiveness of the applied methods.

A methodology based on a combination of analytical procedures and techniques is developed with the main aim of finding an answer to the key question underlying the thesis of the dissertation. The results documented in the dissertation and related publications confirm the existence of efficient ways to quantify and extract valuable information from the studied type of data. Hence, it can be concluded that online chat communication between customers and operators in a contact center represents an unexplored rich source of information about customer relationships with the company. It is demonstrated how this information can be efficiently extracted, structured and analyzed using techniques in the field of natural language processing and machine learning in order to build an automated system for chat data analysis.

The dissertation draws the attention of the research community and the business to the potential applications of the analysis of online chat communication between clients and employees – how such data can be useful for the business, what type of information could be extracted and what techniques for data processing and analysis could be applied in order to extract valuable insights. The most important aspects of such communication (from business point of view) are under focus - main topics of client interest expressed in the discussions with company employees and customer satisfaction with chat services. The last becomes even more important in the context of the COVID-19 pandemic and the huge growth in online communication on a global scale - this also leads to an excessive workload of contact center staff who are at the forefront during the crisis.

The development of such automated analysis system leads to many practical benefits for the business - customer service improvement, better understanding of customer needs and pain points, products and services improvement, identification of trends and topics that concern customers, tracking the effectiveness/success of various campaigns, facilitation of the process of chatbot/automated dialogue system development and much more. Although the empirical study is focused on a specific business case in the banking domain, **the utilized techniques and the author's proposed approach for interpretation of the results** are applicable in all other industries where such type of data is generated as a result of the company's business processes. The conclusions made in the study might be valuable in any business context in which customer orientation is of highest importance representing a key to business development and future success.

IV. Bibliography[11]

[1] S. Gupta and D. Ramachandran, "Emerging market retail: transitioning from a product-centric to a customer-centric approach," Journal of Retailing, vol. 97, no. 4, pp. 597-620, 2021.

[2] M. A. Camilleri, "The use of data-driven technologies for customer-centric marketing," International Journal of Big Data Management, vol. 1, no. 1, pp. 50-63, 2020.

[3] Dimension Data, "2017 Global Customer Experience Benchmarking Report. Digital crisis or redemption. The uncomfortable truth," 2017.

[4] A. Qasem and W. Alhakimi, "The Impact of Service Quality and Communication in Developing Customer Loyalty: The Mediating Effect of Customer Satisfaction," Journal of Social Studies, vol. 25, no. 4, 2019.

[5] B. Lobe, D. Morgan and K. A. Hoffman, "Qualitative data collection in an era of social distancing," International Journal of Qualitative Methods, vol. 19, no. 1-8, Art. no. 1609406920937875, 2020.

[6] KPMG, "Standing firm on shifting sands. Global banking M&A outlook H2 2020," 2020.

[7] Mordor Intelligence, "Big Data Analytics In Banking Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026)," 2021.

[8] E. Pronoza, A. Pronoza and E. Yagunova, "Extraction of Typical Client Requests from Bank Chat Logs," in Mexican International Conference on Artificial Intelligence, 2018.

[9] S. Kumar, A. K. Kar and P. V. Ilavarasan, "Applications of text mining in services management: A systematic literature review," International Journal of Information Management Data Insights, vol. 1, no. 1, Art. No. 100008, 2021.

[10] J. Wang, J. Wang, C. Sun, S. Li, X. Liu, L. Si, Z. M. and G. Zhou, "Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[11] M. Zaki and A. Neely, "Customer experience analytics: dynamic customer-centric model," Handbook of Service Science, Volume II, pp. 207-233, 2019.

[12] S. Roy, R. Mariappan, S. Dandapat, S. Srivastava, S. Galhotra and B. Peddamuthu, "QART: A System for Real-Time Holistic Quality Assurance for Contact Center Dialogues," in Thirtieth AAAI conference on artificial intelligence, 2016.

[13] M. Anandarajan, C. Hill and T. Nolan, Practical text analytics. Maximizing the Value of Text Data, Springer, 2019.

[14] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill and B. Nisbet, Practical text mining and statistical analysis for non-structured text data applications, Academic Press, 2012.

[15] A. Magueresse, V. Carles and E. Heetderks, "Low-resource languages: A review of past work and future challenges," arXiv:2006.07264 [cs.CL], 2020.

[16] K. Park, J. Kim, J. Park, M. Cha, J. Nam, S. Yoon and E. Rhim, "Mining the minds of customers from online chat logs," in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015.

[17] G. Hristova, "Text Analytics in Bulgarian: An Overview and Future Directions," Cybernetics and Information Technologies, vol. 21, no. 3, pp. 3-23, 2021.

[18] G. Hristova, "A SURVEY OF TEXT MINING METHODS APPLIED ON CONVERSATIONAL DATA," Scientific Research of the Union of Scientists in Bulgaria – Plovdiv, series B. Natural Sciences and Humanities, Vol XX. VIIIth International Conference of Young Scientists, 2020.

[19] C. H. Chen, W. P. Lee and J. Y. Huang, "Tracking and recognizing emotions in short text messages from online chatting services," Information Processing & Management, vol. 54, no. 6, pp. 1325-1344, 2018.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, pp. 5998-6008, 2017.

[21] B. Velichkov, I. Koychev and S. Boytcheva, "Deep learning contextual models for prediction of sport event outcome from sportsman's interviews," in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019.

[22] Y. Dinkov, I. Koychev and P. Nakov, "Detecting Toxicity in News Articles: Application to Bulgarian," arXiv:1908.09785 [cs.CL], 2019.

[23] M. Hardalov, I. Koychev and P. Nakov, "Beyond English-Only Reading Comprehension: Experiments in Zero-Shot Multilingual Transfer for Bulgarian," arXiv:1908.01519 [cs.CL], 2019.

[24] B. Velichkov, S. Gerginov, P. Panayotov, S. Vassileva, G. Velchev, I. Koychev and S. Boytcheva, "Automatic ICD-10 codes association to diagnosis: Bulgarian case," in CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics, 2020.

[25] A. Popov, P. Osenova and K. Simov, "Implementing an End-to-End Treebank-Informed Pipeline for Bulgarian," in Proceedings of the 19th Workshop on Treebanks and Linguistic Theories, 2020.

[26] G. Hristova, B. Bogdanova and N. Netov, "Design of ML-based AI System for Mining Public Opinion on E-government Services in Bulgaria," in AIP Conference Proceedings, (accepted for publication).

[27] G. Hristova, B. Bogdanova and N. Netov, "Data Mining of Public Opinion: An Overview," in AIP Conference Proceedings, (accepted for publication).

[28] B. Bogdanova and E. Stancheva-Todorova, "ML-based predictive modelling of stock market returns," in AIP Conference Proceedings, 2021.

[29] G. Mengov, I. Nenov and I. Zinovieva, "A model for collective emotion forecasts financial data," IFAC-PapersOnLine, vol. 52, no. 25, pp. 208-213, 2019.

[30] I. Ivanov, S. Kabaivanov and B. Bogdanova, "Stock market recovery from the 2008 financial crisis: The differences across Europe," Research in International Business and Finance, vol. 37, pp. 360-374, 2016.

[11] The full list of references is available in the dissertation.

[31] E. Stancheva-Todorova and B. Bogdanova, "Enhancing investors' decision-making–An interdisciplinary AI-based case study for accounting students," in AIP Conference Proceedings, 2021.

[32] I. Nenov, G. Mengov, K. Ganev and R. Simeonova–Ganeva, "Neurocomputational economic forecasting with a handful of data," Comptes rendus de l'Académie bulgare des Sciences, vol. 74, no. 10, 2021.

[33] B. Kapukaranov and P. Nakov, "Fine-grained sentiment analysis for movie reviews in Bulgarian," in Proceedings of the International Conference Recent Advances in Natural Language Processing, 2015.

[34] J. Boyd-Graber, Y. Hu and D. Mimno, "Applications of Topic Models," Foundations and Trends in Information Retrieval, vol. 11, no. 2-3, p. 143–296, 2017.

[35] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, vol. 78, no. 11, pp. 15169-15211, 2019.

[36] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," EAI Endorsed transactions on scalable information systems, vol. 7, no. 24, 2020.

[37] A. M. Dean, "The Impact of the Customer Orientation of Call Center Employees on Customers' Affective Commitment and Loyalty," Journal of Service Research, vol. 10, no. 2, pp. 161-173, 2007.

[38] A. Rafaeli, L. Ziklik and L. Doucet, "The Impact of Call Center Employees' Customer Orientation Behaviors on Service Quality," Journal of Service Research, vol. 10, no. 3, pp. 239-255, 2008.

[39] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[40] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1-167, 2012.

[41] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams engineering journal, vol. 5, no. 4, pp. 1093-1113, 2014.

[42] S. Baccianella, A. Esuli and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), 2010.

[43] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," arXiv:1103.2903 [cs.IR], 2011.

[44] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in Proceedings of the International AAAI Conference on Web and Social Media, 2014.

[45] W. L. Hamilton, K. Clark, J. Leskovec and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in Proceedings of the conference on empirical methods in natural language processing, 2016.

[46] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowledge-Based Systems, vol. 89, pp. 14-46, 2015.

[47] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang and C. Liu, "A survey on deep transfer learning," in International conference on artificial neural networks, 2018.

[48] G. Hristova, "Topic modeling of chat data: A case study in the banking domain," in AIP Conference Proceedings, 2021.

[49] G. Hristova, "Topic Modeling of Chat Data: Experimenting with Different Levels of Chat Data Representation by Utilizing a Latent Dirichlet Allocation Model," Journal of Economic Boundaries And Transformation, (accepted for publication).

[50] G. Hristova, "Text Analytics for Customer Satisfaction Prediction: A Case Study in the Banking Domain," in AIP Conference Proceedings, (accepted for publication).

[51] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017.

[52] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[53] M. Röder, A. Both and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015.

[54] P. Nakov, "BulStem: Design and evaluation of inflectional stemmer for Bulgarian," in Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics), 2003.

[55] K. Simov, P. Osenova and M. Slavcheva, "BTB-TR03: BulTree-Bank Morphosyntactic Tagset. BTB-TS version 2.0.," Technical report, Bulgarian Academy of Sciences, Sofia, Bulgaria, 2004.

[56] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, New York: Springer, 2013.

[57] H. Lei and Y. Chen, "Concentrated Document Topic Model," arXiv:2102.04449 [stat.ML], 2021.

[58] R. Taylor and J. A. D. Preez, "ALBU: An approximate Loopy Belief message passing algorithm for LDA to improve performance on small data sets," arXiv:2110.00635 [cs.LG].

[59] K. Stevens, P. Kegelmeyer, D. Andrzejewski and D. Buttler, "Exploring topic coherence over many models and many topics," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012.

[60] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014.

[61] J. Qiang, Z. Qian, Y. Li, Y. Yuan and X. Wu, "Short text topic modeling techniques, applications, and performance: a survey," IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 8, 2020.

V. Contributions of the dissertation

1. Development of an extensive literature resource presenting an up-to-date picture of the development in the field of natural language processing in Bulgaria, as well as the opportunities for analyzing textual data in Bulgarian.
2. Development of a research method for interpretation of the obtained results, characterized by applicability and possible extrapolation of its steps. The approach is applicable not only in the banking industry, but also in any other domain in which similar type of data is generated as a result of the company's business processes.
3. Development of an automated approach for knowledge extraction from online chat communication with the customer conducted in a contact center. The illustrated methods for quantitative analysis could be combined with established techniques for qualitative analysis so that the business companies could take better decisions characterized by a higher degree of rationality.
4. The possibility to adapt the developed automated system to new data and to update the information in real time could be considered an additional contribution of the dissertation. It is important to note that although the system is adaptive, there is a need for periodic monitoring and update of the prediction models with regard to the changing characteristics of the data over time.
5. Development of a methodology for the analysis of online chat communication with customers conducted in the Bulgarian language. Specific algorithms and procedures for data processing and analysis could be replicated by other researchers in the field and used to automate and facilitate the analysis of similar data (e.g., creation of libraries, modules, etc.).
6. Development of a research method for predicting customer satisfaction with online chat communication, which is based solely on textual features combined with grammatical information extracted from textual data.
7. Development of a research method for analyzing the topics of client interest. Important text processing and modeling techniques aimed at knowledge extraction from the studied type of data, are illustrated. Valuable conclusions are drawn regarding the different levels of chat data representation - a topic which, to the best of the author's knowledge, has not been addressed by other researchers in the field until now.

VI. List of publications related to the dissertation

1. G. Hristova, "A SURVEY OF TEXT MINING METHODS APPLIED ON CONVERSATIONAL DATA," Scientific Research of the Union of Scientists in Bulgaria – Plovdiv, series B. Natural Sciences and Humanities, Vol XX. VIIIth International Conference of Young Scientists, 2020.
2. G. Hristova, "Text Analytics in Bulgarian: An Overview and Future Directions," Cybernetics and Information Technologies, vol. 21, no. 3, pp. 3-23, 2021.
3. G. Hristova, "Topic modeling of chat data: A case study in the banking domain," in AIP Conference Proceedings, 2021.
4. G. Hristova, "Topic Modeling of Chat Data: Experimenting with Different Levels of Chat Data Representation by Utilizing a Latent Dirichlet Allocation Model," Journal of Economic Boundaries and Transformation, (accepted for publication).
5. G. Hristova, "Text Analytics for Customer Satisfaction Prediction: A Case Study in the Banking Domain," in AIP Conference Proceedings, (accepted for publication).