R E F E R E N C E

From: Prof. DSc Zhelyu Vladimirov, professional field 3.7 "Administration and management"

To: dissertation work by Gloria Ventsislavova Hristova, on the topic "*Automated system for the analysis of online communication with customers through machine self-learning and natural language processing - structure, construction and business applications*", for awarding the educational and scientific degree "Doctor" in professional field 3.8 "Economics", scientific specialty "Data analytical research".

Reason for the review: Order RD-38-326/04.07.2022 of the Rector of Sofia University "St. Kliment Ohridski"

1. Information about the doctoral student

Gloria Ventsislavova Hristova graduated from Sofia Mathematical High School (SMG) "Paisiy Hilendarski" (2004-2012). In 2016, she graduated from the Bachelor's program "Business Administration" at the Faculty of Economics at the Sofia University "St. Kliment Ohridski". In 2018, she defended his master's thesis on the topic "Design and development of an automated system for determining user attitudes towards software applications using natural language processing techniques and machine self-learning" and graduated with honours from the master's program "Big Data Modelling in Business and Finance". Gloria Hristova continues her studies at the Faculty of Economics as a full-time doctoral student in the "Data Science" doctoral program.

From September 2021, Gloria Hristova holds the position of "assistant" at the Department of "Statistics and Econometrics" at the Faculty of Economics, where she leads classes on "Quantitative methods in management", "Machine learning for business and finance" and "Basics of text analysis and processing in natural language". In the last three years, she took part in four national scientific projects, as well as in several seminars and international conferences. She also participated as a mentor in competitions in the field of analytical research.

2. General characteristics of the presented dissertation work

The presented dissertation consists of an introduction, three chapters, a bibliography and appendices with a volume of 281 pages, and without the appendices and information sources – of 252 pages. A total of 220 information sources in English are used. The main text contains 26 tables and 27 figures, with a further 5 tables and 10 figures given in the appendices. According to the reference for fulfilling the criteria, doctoral student Gloria Hristova has 320 points with a required minimum of 150 points.

In the **Introduction**, the relevance of the issues related to the spread of "big data" and the need for its storage, processing and use by companies is revealed. It is pointed out that data extraction, processing and interpretation acquire a central role in the conditions of online communication, since such data reflects the main problems that concern customers (p. 12).

In this regard, the main **goal** of the research is defined as: "Creating an automated system for the analysis of the main topics that excite customers, as well as for the analysis of their satisfaction with the services provided in a contact centre with communication in the Bulgarian language" (p. 16). This objective is specified in three major tasks with eight sub-tasks. Its achievement is based on the use of various analytical techniques from the field of natural language processing and machine learning. The **object** of the dissertation is the online chat communication between customers and employees of a large bank in Bulgaria, while the chosen **perspective** are

the main topics that excite customers, as well as their satisfaction with communication through chat with the contact centre. The data is generated in the contact centre of this bank.

The thesis defended is that online chat communication between customers and contact centre operators can be effectively extracted, structured and analysed using natural language processing and machine learning techniques by building an automated analysis system (p. 19). The verification of the thesis is realized by testing 4 hypotheses.

3. Evaluation of the obtained scientific and scientific-applied results

Chapter one includes a detailed and critical literature review of current research since 2016 on the analysis and knowledge extraction of online chat communication with a focus on customer service in a contact centre. In point 1.2. the emphasis is on: describing this communication, modelling topics, and solving classification tasks. Table 2 (pp. 58-59) presents a summary of methods for extracting knowledge from online chat communication. It is concluded that no studies were found analyzing online chat communication in Bulgarian (p. 61).

An up-to-date picture of natural language processing and the application of analytical techniques on textual data in Bulgarian is given in point 1.3. Research on the creation of language resources in Bulgarian, mainly by the BulTreeBank group, is presented. Tools and systems for processing text in Bulgarian are shown (tokenization, stemming, lemmatization, etc.). Practical applications of textual data analysis in Bulgarian for solving social, economic and business problems are revealed (Table 3, pp. 79-80). The stages in natural language processing are outlined – from the use of rules through statistical methods and machine learning to transfer learning.

Section 1.4 presents methods for modelling topics from textual data in online chat communication. The author chose the LDA (Latent Dirichlet Distribution) algorithm to model and analyse the underlying themes. The main approaches to sentiment analysis from textual data (the use of lexicons, machine learning, or a combination of the two) are disclosed in *Section 1.5*. Machine learning methods and logistic regression are chosen to predict customer satisfaction with online chat communication.

The second chapter includes a detailed exposition of the research methodology. The creation of 4 modules that make up the automated system for analyzing customer communication with a contact centre is explained. Module I contain an algorithm for initially reading and structuring data from online chat communication using the Python programming language. The stages for converting the data from a raw form to a form suitable for processing and quantitative analysis are shown. The general steps in the algorithm for transformation and normalization of textual data are presented in Fig. 4 (p. 125).

The proposed methodology in Module II represents an author's combination of different approaches and techniques for modelling topics in online chat communication. The basic techniques for normalizing text before converting it to numeric form are given in Fig. 5 (p. 130). After data normalization, word-level "tokenization" was applied to the data, resulting in each utterance from a given chat being represented as a set of tokens (string of words). Levels of text representation are indicated depending on whether whole chats or only some lines are used. To extract the main topics in online chat communication, TF-IDF vectorization of the through the "gensim" library in Python is used. The choice of LDA is justified, which considers each document as a probability distribution of the set of all themes (Fig. 6, p. 140). Two metrics are described for evaluating the results of LDA application (complexity ratio and coherence metrics - p. 145).

Module III presents the author's automated method for predicting customer satisfaction with contact centre services. Three levels of data representation are created: a sample of entire customer-operator chats (“Sample 1”); a sample including all customer-only replicas (“Sample 2”); and a

sample including only the client's final lines in the communication ("Sample 3"). Three different algorithms are tested - the Bernoulli naive Bayesian model, logistic regression and support vector classification. The author uses a $\boldsymbol{k}$-fold cross-validation to evaluate the models, as well as precision, sensitivity, F1 and F-beta against the "poor rating" class with a higher model sensitivity weight.

A summary of the results are presented in *Module IV*, which aims to facilitate their interpretation. As the author writes, while Modules I, II and III represent the "back-end" of the system, Module IV is its "front-end" - i.e. what the end user interacts with directly (p. 180).

In the **third chapter**, the automated system is tested by analyzing the online chat communication between customers and operators in a contact centre of a large financial institution in Bulgaria. The data sample consists of 38,166 chats, the period is from 22/01/2019 to 01/04/2021, and the total number of replicas is 466,118. The final number of chats after data normalization in Module I is 37,529 observations. The object of the analysis are chats with a customer and one operator, which are 29,614 or 78.9% with an almost even distribution of replies. Data on average chat duration, average response speed, and more are presented. The analysis of the main topics of interest to customers is carried out using the LDA algorithm. The two quantifiers are used to select the optimal number of topics (Cv coherence and complexity ratio).

A total of 1,470 models are created to estimate the mean Cv coherence for models with 2 to 50 subjects. It is concluded that between 15 and 20 topics, the coherence of the model reaches a stable average value - 16 themes after the fifth filtration and with the smallest number of words (Table 15, p. 200). For the fifth filtration, the value of the complexity coefficient is also calculated, and again the analysis shows a decrease in the value between 14-16 topics (Fig. 23, p. 203).

Point 3.3.1 analyses the created model with optimal performance on Sample I of each topic separately. Figure 24 (p. 212) demonstrates four very common themes in customer discussions: 1. Lending, 2. Digital banking, 3. Cash operations, and 4. Card products (credit and debit cards). A similar analysis is made on Sample II and III. It is concluded that the highest quality of themes is achieved by modelling the entire chats (Sample I), which leads to the rejection of Hypothesis 2. At the same time, the optimal results obtained on Sample I lead to the confirmation of Hypothesis 1. This means that the cues of the communication operators are necessary to extract the important for customers themes.

In pint 3.6. it is shown how, using previous data and machine learning, characteristics of customer-agent communication can be captured with sufficient accuracy to signal customer satisfaction or dissatisfaction with the contact centre service (Hypothesis 3 and Hypothesis 4). Three different representations of the data (Sample 1, 2, and 3) are used, creating 216 models predicting this satisfaction. According to the results, the chats themselves contain enough signals to create a model predicting the customer's satisfaction at the end of the communication, which leads to the confirmation of Hypothesis 3. Also, the customers' final remarks contain valuable information about their mood at the end of the communication according to the F-beta metric, which leads to the confirmation of Hypothesis 4.

The **conclusion** represents a recapitulation of the conducted research, and some new perspectives in this area are indicated.

4. Evaluation of scientific and scientific-applied contributions

Several contributions are formulated, which can be summarized as follows: (1) An up-to-date picture is presented in the field of natural language processing in Bulgaria, as well as the possibilities for analyzing textual data in Bulgarian; (2) A comprehensive methodology for processing and analyzing data from online chat communication with clients in Bulgarian is proposed; (3) An automated system was created to extract knowledge from online chat

communication with customers in a contact centre, which could be updated in real time as new data is received; (4) A methodology for the analysis of important to clients themes is constructed, including basic techniques for data processing and modelling with different levels of representation; (5) A methodology is created for predicting customer satisfaction with online chat communication, which is based only on textual characteristics and grammatical information; (6) A methodology for the interpretation of the obtained results is developed, which can also be used in other industries where similar data are generated; (7) The present study is among the first to analyse and extract knowledge from online chat communication between customers and employees in Bulgarian.

5. Evaluation of dissertation publications

The main components of the dissertation work are tested in five publications, of which three in international conferences and two in journals in the period 2020-2021. Three of the publications are indexed in globally recognized databases (Scopus and Web of Science).

6. Evaluation of the dissertation summary (autoreferat)

The summary of the dissertation is 54 pages long. It reveals the main points of the dissertation work in a synthesized form and as such meets the requirements.

7. Critical notes, recommendations and questions

I have no specific notes on the content of the text. Only from an editorial point of view, the repetition of the hypotheses on pp. 101, 102, 112 and 114 seems redundant.

8. Conclusion

The doctoral student has done significant work on analyzing the relevant literature, conducting empirical experiments, testing hypotheses using adequate metrics for evaluating the results obtained and the applied methods. The issues are relevant and the conclusions drawn add value to the available knowledge on the subject. A very good knowledge of the researched issues, accuracy in the presentation of the theoretical approaches, correctness in referring to the used publications are demonstrated. The results confirm the existence of effective ways to analyse and extract information from online customer communication using natural language processing and machine learning techniques. An automated system is proposed for the analysis of the main topics that excite customers, as well as for the analysis of their satisfaction with the services provided in a contact centre with communication in the Bulgarian language. The practical benefits for business of creating such a system are substantiated. The contributions made are the personal achievements of the author. With this work, Gloria Hristova demonstrates the qualities of a serious researcher and a responsible attitude to scientific activity

All this gives me a reason to propose to the respected scientific jury to award Gloria Ventsislavova Hristova the educational and scientific degree "Doctor" in professional field 3.8 "Economics", scientific specialty "Data analytical research".

10/08/2022
Sofia

Reviewer:
Prof. DSc Zhelyu Vladimirov