Approach for Social Media Content-Based Analysis for Vacation Resorts

Snezhana Sulova, and Boris Bankov

Original scientific paper

***Abstract*—The impact of social networks on our lives keeps increasing because they provide content, generated and controlled by users, that is constantly evolving. They aid us in spreading news, statements, ideas and comments very quickly. Social platforms are currently one of the richest sources of customer feedback on a variety of topics. A topic that is frequently discussed is the resort and holiday villages and the tourist services offered there. Customer comments are valuable to both travel planners and tour operators. The accumulation of opinions in the web space is a prerequisite for using and applying appropriate tools for their computer processing and for extracting useful knowledge from them. While working with unstructured data, such as social media messages, there isn't a universal text processing algorithm because each social network and its resources have their own characteristics. In this article, we propose a new approach for an automated analysis of a static set of historical data of user messages about holiday and vacation resorts, published on Twitter. The approach is based on natural language processing techniques and the application of machine learning methods. The experiments are conducted using software product RapidMiner.**

***Index Terms*—Text Mining, Support Vector Machines, Machine learning, Twitter.**

I. Introduction

In the last couple of years, we have witnessed an increased interest towards the opportunities of data mining of social networks. Websites such as Facebook, Twitter and Reddit are rich with text data, as well as links to other resources such as images, videos and users' activities data [1]. It has never been this difficult to analyze and summarize unstructured data. Social networks are thriving [2]. Exchanging information happens instantaneously. But this also poses a challenge – business decisions need to be synchronized according to customer questions, reviews and expectations, formulated on the basis of user opinions.

There are different works in the field of social media mining. Some are using visual cluster projections of user interests to map out relationship networks [3], [4]. Others use RapidMiner and neural networks to evaluate the effect and performance of paid advertisements on websites like Facebook [5].

Manuscript received January 27, 2019; revised September 5, 2019. Date of publication September 18, 2019. Date of current version September 18, 2019.

Authors are with University of Economics – Varna, Department of Informatics, Bulgaria (e-mails: ssulova@ue-varna.bg, boris.bankov@ue-varna.bg).

Digital Object Identifier (DOI): 10.24138/jcomss.v15i3.712

In this article, we focus our attention on the effect of the developing information technologies in travel and tourism business environment. Companies use the Internet to promote and present information about their products and services. Alongside that customers have the opportunity to purchase or make a reservation online as well as leave an impression or a comment about their experiences. More and more customers begin to rely on social networks to publicly express their opinions. Roughly 6000 tweets are being published worldwide every second [6]. This of course means Twitter is one of the biggest and richest streams of valuable information that attracts research in the field of data mining and textual analysis.

In tourism, any form of personal user review can be used by resort and hotel managers to quickly receive information about how their customers feel and how their business is perceived. On the other hand, customer opinions can be easily shared online and they often have significant impact on other users' decisions in regard to their own vacation plans.

We can conclude that there are enough prerequisites to justify the growing amount of research being done to solve the problems of textual data mining and analysis of unstructured data, originating from social networking web sites.

With the following article, we propose an approach for extracting data from Twitter to complement automated sentiment analysis of text data. To test our approach, we chose to look at mentions of Croatian holiday and vacation resorts in public tweets.

In general data analysis papers may skip or present less focus on the techniques of extracting the sample data or the pre-processing stage. Twitter allows for multiple approaches. Our goal is to be able to look at a massive array of data of static tweets, without any knowledge of the context in which each user message appears. Then filter tweets based on a predefined set of rules in order to cleanse the data, grab specific messages that contain mentions about Croatian tourist destinations and extract the polarities of opinions in those selected posts for the 2016, 2017 and 2018 month of July.

Our goal is to look at data regarding the same period of time during different years when tweets occurred. To accomplish this, we need an algorithm which can retrieve and pre-process tweets with a time constraint. While Twitter allows subscribing and listening to public streams based on predefined filters in real time, for any researcher that would mean that they had to have already been listening to that data back in 2016, 2017 and in 2018.

This can be achieved if a computer configuration is running 24/7 for the whole month of July during all three years and only after the process is complete researchers can look at the gathered data. Although this is not impossible to do, it is far more plausible in our case to conduct a search that doesn't take two years to accomplish. Twitter's Search API is very limited, and it is not suited for such experiments. Our goal is to prepare an algorithm which can be fully automated and would be a solution for when researchers need to extract data from a static set of past tweets.

The rest of the paper is organized as follows. Section II presents an overview of the theoretical basis and methodologies used in text mining. Sentiment Analysis is discussed and in particular machine learning algorithms and models such as SVM and Naïve Bayes. Section III deals with the different sides of Twitter's API technological possibilities for retrieving data from the platform. This section introduces a list, created and compiled by the authors, consisting of regular expression filters that are best suited for sanitizing and cleansing tweets in the data retrieval stage. Section IV is divided in two parts. The first part shows the process we have developed for extracting Twitter messages from a static collection of historical data for the month July in 2016, 2017 and 2018. The second part is where the classifiers are applied and the data model is trained as a means to exercise and validate the usefulness of the data. In the final section the results from the classification process are discussed.

II. Theoretical Foundations and Technologies for Text Mining

The increase of unstructured data, which according to analysts accounts for more than 80% of all data [7], is becoming a prerequisite for the development of technologies for text analysis. In general, knowledge retrieval and extraction from unstructured text data is known as Text Mining (TM) [8], [9], [10]. In modern literature, the term TM is also known as Text Data Mining (TDM), Text Analytics, Knowledge-Discovery in Text (KDT) [7].

The process of text mining can include the following:

1) Information retrieval (IR) – best described as searching and finding relevant information in databases, in document repositories, web resources on the Internet as well as extraction of document meta-data.

2) Natural language processing (NLP) – it involves text analysis, transformation and formal presentation of language resources in a format, appropriate and allowing for machine computations;

3) Information extraction (IE) – relates to the activity of extracting structured information from unstructured environments.

4) Data mining (DM) – by looking for patterns and behavior within data, the goal is to discover hidden relations and knowledge [11].

Pena-Ayala shares a research that summarizes the typical TM approaches as follows: text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling [12].

The rapid increase of information volumes on the Internet undoubtedly gives plenty of reasons for data scientists to look into ways of retrieving valuable data from web resources such as web pages. Modern literature has coined the term Web Mining (WM). One of the first pioneers in the field is Etzioni, who describes WM as the "use of data mining techniques to automatically discover and extract information from World Wide Web documents and services (e.g., on-line travel agents, job listings, electronic malls, etc.)" [13]. In scientific research, 3 types of VM are defined depending on the resources analyzed [14]: 1) Web Content Mining (WCM) – the source of data is the content of the web pages; 2) Web Structured Mining (WSM) – extracting useful knowledge from the structure of web sites; Web Usage Mining (WUM) – extracting useful knowledge from data on the use of Internet resources.

Many researchers have published works about WCM [15], [16], [17], [18]. Their studies aim to solve different problems. Materna, Qi and Davison work on automated web page classification [19], [20]; Markov and Larosh focus on text document grouping [18]; Hariharan, Srinivasan and Lakshmi – similarity discovery of text documents [21], [22]; Liu, Medhat, Hassan, Korashy, D'Avanzo, Pilato, Patel, Prabhu and Bhowmick conduct research in two areas: Sentiment Analysis and Opinion Mining [23] ,[24], [25], [26], [27].

In conclusion, we can safely state that any and all kinds of text analysis require the use of DM methodology. The difference in how the problems are approached is due to the nature of unstructured data and the various types of text resources. In our research, we focus on Sentiment Analysis (SA) and Opinion Mining (OM), fields of research dealing with automated extraction of subjective human point of view – his opinion, emotional view and attitude towards a specific topic, event, product or service.

Opinion Mining as a term is introduced by Dave, Lawrence and Pennock. The definition is as follows: "a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)" [28]. In literature we can find other, broader definitions and interpretations of opinion mining. Sadegh, Ibrahim and Othman describe opinion mining as a selection of techniques for discovery and retrieval of subjective information in text documents [29]. Opinion Mining is based on natural language processing technologies and the focal point lies in the extraction of perceptions, opinions, views and ideas, while the automated analysis targets attribute retrieval and whether the analyzed text expresses a positive, negative or neutral sentiment.

Sentiment Analysis is first mentioned in the works of Das, Chen and Tong [30], [31]. They use it alongside automated analysis and text evaluation. What follows are numerous research papers that deal with different aspects of Sentiment Analysis and in many of them the term is used as a synonym for opinion mining, due to the fact that sentiments are retrieved primarily from documented opinions [29], [32], [33], [34].

Sentiment Analysis applies to subjective information as it aims to produce conclusions based on natural language processing, computer linguistics, statistical methodologies and

artificial intelligence. In his continued work Liu studies SA and differentiates approaches depending whether it is a word, sentence or a whole document that is being analyzed. When a document is being analyzed the resulting conclusion is generalized. The sentiment can be positive, negative or neutral, but this label is placed on the whole document, individual polarity of text segments cannot be deduced. The approach is similar for analyzing a sentence – the outcome can be positive, negative or neutral. When single words are being looked at, special taxonomies are built, and it is crucial to note that not always a word with significant emotional weight can determine the sentiment outcome for the whole sentence or text [25].

Research in OM and SA suggests the application of various methodologies, mainly focused around machine learning and creating taxonomies or a combination of both [26], [35], [36]. Figure 1 shows a summary of SA methods [37]. They are separated into three major categories: machine learning, language collections and hybrid approaches (Fig 1).

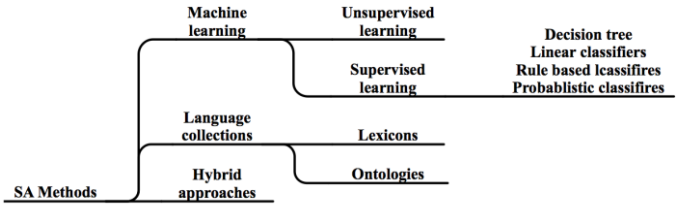

Fig. 1. Characteristics of DM, TM and WM.

When analyzing text arrays various taxonomies and language dictionaries are used [38], [39]. Building a general-purpose taxonomy is difficult because different problems require particular vocabularies. Such examples are the English corpus WordNet, Croatian CroWN or the Bulgarian BulNet.

Some methodologies are based on ontologies, which contain a formal description about the subject, finite vocabularies of terminologies and eloquent examples of relationships in specific context. There are also SA hybrid approaches that combine language collections and machine learning.

Methods based on machine learning are divided in two groups: Supervised Machine Learning – applicable and better suited for text analysis of labeled datasets – and Unsupervised Machine Learning, which is used often when there are no labels and the text data can be unpredictable. Most used sentiment classification algorithms are Support-Vector Machines (SVM) and Naïve Bayes [40], [41].

SVM is a binary linear classifier that uses training examples and maps them to points in n-dimensional space. The algorithm tries to solve the mapping so a hyperplane could be created, where data is divided into two classes: a set of positive and a set of negative examples with maximum margin (Fig. 2). The linear classification function f(x) is expressed as:

$$f(x) = w^t x + b \quad (1)$$

where: w^t is the normal to the line or the weight vector and b is the bias.

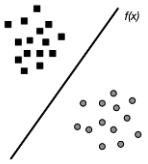

Fig. 2. Linear Classifier SVM

Naïve Bayes (NB) is a probabilistic classifier. The mechanism behind this algorithm consists of computing the conditional probability that an object is part of specific class based on known features about the data. The algorithm relies on the application of a "naive" assumption that there is independence between every pair of traits given the value of the class variable. Given $P(y= c_r)$ is the probability that an object i_j is part of the class c_r, the event is E, and the likelihood of which is $P(E)$, then Bayes' theorem is:

$$P(y = c_r|E) = \frac{P(E|y=c_r)\,P(y=c_r)}{P(E)} \quad (2)$$

It should be noted that sentiment analysis and text analysis are essentially a non-traditional task of processing unstructured text data. They do not have a universal algorithm and a specific rule as to which methods and technologies are recommended to be used in each case. When retrieving data from Internet resources, such as social networks, specific ways of accessing and retrieving content are used. In this article we offer an approach for retrieving, processing and sentiment analysis of data from the social network Twitter.

III. Approach for Analysis of User Mentions from Twitter

It is important to note that in scientific literature we can already find a number of different approaches for opinion mining from web sites and social media platforms.

Lai and To suggest an algorithm based on computer lexical analysis, accompanied with statistical and graphical methods, which help identifying and categorizing key terms [42]. They use Scott lexical software [43] and lexical mapping software Leximancer. The approach is more suited for analysis of customer opinions, for example on a web site for a hotel or resort rather than social media.

Saggion and Funk [44] apply a different methodology. Their approach is based on lexical resource SentiWordNet and SVM classification. However, they do not offer a tool for retrieval and pre-processing for specific text arrays such as the ones found on social network web sites.

Research focused on social media mining can be found in the works of authors [45] which apply a personalized social search based on the user's social relations. Others [46], [47] offer location discovery based on an analysis of user`s post content in social platforms. Both these specialized approaches work for particular scope of interest.

Based on the examination of the state of opinion mining of social media platforms we believe analyzing user`s comments can be improved by applying our proposed approach on top of other researcher`s methodologies. The retrieval and pre-processing of tweets from Twitter can be adjusted and modified

to work with different data sources or social platforms and can be suited for the needs of various studies in economy and marketing or other fields.

To apply Opinion Mining in an unstructured environment such as a collection of tweets we propose an approach, consisting of the following steps (Fig. 3):

1) Data retrieval – collecting comments, questions, mentions on Twitter about specific resorts.

2) Text processing – data cleansing, removal of insignificant words and creating vectors from words (word embeddings).

3) Data mining – applying algorithms for building predictive models for SA and OM.

4) Results interpretation – analysis and evaluation of output.

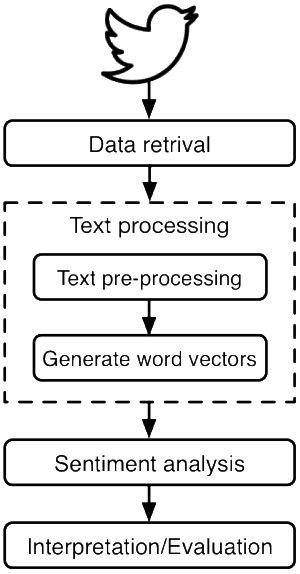

Fig. 3. Stages of Twitter data retrieval and processing

A. Twitter Data Retrieval

Twitter data streams are a rich source of various information about holiday and vacation resorts. Any Twitter message (called a tweet), can contain up to 280 symbols long opinion or a fact about any given matter. Apart from the text body of a tweet, usually one message contains nearly 140 additional attributes, such as the timestamp, information about the author, the location, where the tweet originates, number of shares and likes and so on. Based on a review of about 90+ software applications that improve the quality of life of researchers and developers when working with Twitter, we can categorize the tools as follows:

1) Tools for marketing analysis.

2) Tools for following discussions/chat.

3) Tools for finding the latest information and new users.

4) Tools for hashtag processing and analysis.

5) Tools for notifications and monitoring [48].

There are two ways of retrieving a significant amount of data for research purposes from Twitter. The first is to get a static collection of past tweets – this however is a premium service that costs companies like Microsoft, Apple and Google millions of dollars. The other way to download data straight from Twitter is via a public stream. To access that stream up until 2017 there were three application programming interfaces (APIs):

1) Twitter's Search API.

2) Twitter's Streaming AP.

3) Twitter Firehose.

Twitter's Search API allows access to a predefined collection of tweets, which have been stored for some time. The request for data is initiated by the user and approximately 1% of all tweets can be viewed in the matching results. *Twitter's Streaming API* offers a direct connection to a live stream of tweets, that are currently being posted on the platform. The request for data is initiated by Twitter following specific criteria, given by the user and the result is anywhere between 1% and 40% of all real-time tweets. *Twitter Firehose* is a paid service that corporations can use to retrieve a sizable amount of information, initiated by Twitter. Firehose guarantees 100% data integrity.

After changes made to Twitter's developer API in 2017, the platform introduced an enterprise solution in the face of PowerTrack API, which supports more filtering options and connections than Track API – the standard/free real-time streaming interface.

For the purposes of our research we choose to use a public archive of tweets from the general twitter stream [49]. The technology used to catch the stream is called "Spitzer" which is layered on top of a paid Firehose stream. The Spitzer stream catches between 4% and 15% of all tweets. *We chose to retrieve all archived tweets for the month of July 2016, 2017 and 2018, as during that period we expect a concentration of holidays and vacations to occur and as such more mentions of resorts in tweets and look for messages that contain the names "Dubrovnik", "Rijeka" and "Plitvice".*

B. Text pre-processing for Sentiment Analysis

When working with raw text data it is important to properly supply the input so that the output is not ambiguous, and it can be used to draw clear conclusions from the results. Data cleansing or data pre-processing is usually applied to text before any other methodology or approach is used, e.g. classification or clusterization. Informal text collections such as chat logs, social media messages, forum posts and so on, often contain a significant amount of words that are not meaningful enough to contribute any knowledge about the analyzed subject.

We chose to handle text processing on two stages. First, a filter is used to cleanse twitter messages based on an algorithm for applying regular expressions. Due to the nature of the mixed formal and informal language in social posts the following steps are ordered in way that gives the best state of tweets after pre-processing. The regex used is written in the server-scripting language PHP.

1) Removing tweets which are automatically created by mobile apps such as: "I liked x video on YouTube" or "I shared x video on YouTube";
2) Removing special retweet symbols: "RT :" – regex (RT .*?:);
3) Removing user mentions: @user – regex (\s*@(.+?)\s);
4) Removing hyperlinks: regex (http(.*)\b);

5) Identifying and removing words which use characters not part of the Latin or the Cyrillic alphabets: regex ([\p{Z}]{2, }/u);
6) Removing common emojis;
7) Removing numbers: regex ([0-9]+);
8) Removing punctuation according to POSIX: ([:punct:]);
9) Removing words that are shorter than 2 characters (if the message is in Unicode format then 4 characters limit during this step).

After the filtering is completed, frequency list is created of all words that are found in tweets, mentioning either Croatian holiday location. The goal is to look at the most frequent and least used words and to decide if the classification can be improved by omitting unnecessary terms such as "the" or "of" or names of people or organizations. Some other helpful techniques that can be used are tokenization and stemming.

C. Application of Data Mining Methodologies for Sentiment Analysis

The majority of related work on Opinion Mining and Sentiment Analysis that is mentioned earlier helped us choose an approach with supervised machine learning, combining two methodologies for text categorization – linear classifier SVM and NB. As shown on Fig. 4 our approach is to do a two-step classification.

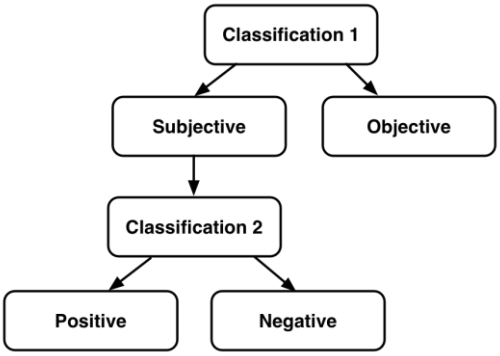

Fig. 4. Classification stages.

During the first stage opinions are split into two categories: subjective (opinions) and objective (facts). The following step shows that subjective statements can be either positive or negative, meaning if they are classified as subjective, they are based on an emotion or a mood.

Calculating the subjectivity of the text segment before retrieving its polarity is a subject of research in [50], [51]. It has been proven that this methodology helps improve the outcome. Emotional polarity recognition is better and improved if the processed data originates only from subjective or opinioned text and all objective or factual statements are removed on the pre-processing stage.

IV. Conducting the Analysis of Twitter Users' Opinions of Three Croatian Tourist Destinations

A. Twitter Data Extraction and Processing

We chose to look at tweets from July 2016, 2017 and 2018, during which we hope to find valuable information about Croatian tourist destinations. We accessed archives from Twitter's public data streams. Each archive contained 31 folders for each day of the month, in each "day" folder there were 24 subfolders for every hour of the day, and within those "hour" subfolders there were 60 archives for each minute. We present the folder structure (as shown on Fig. 5) in order to make it easier to follow our algorithmic approach towards extracting the information.

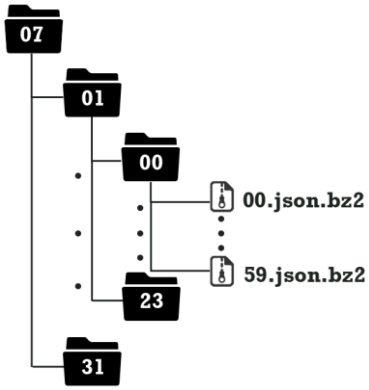

Fig. 5. Folder structure of Twitter archive.

In total there are 744 different directories (31_x24) that contain 44640 archives (31_x24_x60) for July. Each archive holds a JSON (JavaScript object notation) file with tweets. To automate the process of extracting those files on the Windows OS ecosystem Windows PowerShell and WinRaR can be used. The following script allows for fully automated extraction of data from the Twitter archive.

```
1>  $Zips = Get-ChildItem -filter "*.bz2" -path
    "D:\twitter\07\01\" -Recurse
2>  $WinRar = "C:\Program Files\WinRAR\winrar.exe"
3>  $Destination = "D:\twitter\07\01\";
4>  foreach ($zip in $Zips)
5>  {
6>  $dir = $zip.Directory.Name
7>  &$Winrar x $zip.FullName $Destination$dir"\"
8>  Get-Process winrar | Wait-Process
9>  }
```

Line 1 goes through all the contents of the specific folder filtering only archives of type .bz2 and saving them to a variable. Line 2 points to the path of the unpacking software. Line 3 shows the output destination where the files should be unpacked. The final piece of code is a loop that runs the software with the proper parameters.

The next step is to grab and filter the mentions in tweets for Croatian holiday and vacation resorts. One of the fastest ways to skim through the data for the examined period is to use jq - a lightweight and flexible command-line JSON processor. It runs without dependencies from other software, it can be executed from Windows PowerShell and it is able to filter, map and transform JSON files very fast.

As a part of our research we needed to filter and extract only tweets that contained mentions of specific names of Croatian

resorts or famous places of interest for tourists. To do so the tweet's attribute called `text` needs to be convert it to a type string and checked if it contains the name of the resort. Then the algorithm extracts the full message, skips any empty tweets, points to the original JSON file and outputs with append to a csv file (expanding data size could require transferring JSON files to MongoDb instance [52]). As stated earlier the places, chosen for the purpose of this experiment, are Dubrovnik, Rijeka and Plitvice.

```
1>   jq -r 'select(.text | tostring | contains(\"Dubrovnik\")) |
[.text] | select(length > 0) | @csv'
"D:\twitter\07\01\00\00.json" >>
"D:\twitter\results\Dubrovnik.csv"
```

The above example processes data from the first minute of the first hour (midnight at 00:00) of the first day of July 2016 (compare the file path and the folder structure on Fig. 5 for further clarification). To repeat and execute the process for multiple files, multiple filters and output to different files PowerShell code is used to wrap the jq script in a loop to recursively look through folders and files.

```
1>  $files = Get-ChildItem "D:\twitter\07\01\" -Recurse -
    Include *.json
2>  for ($i=0; $i -lt $files.Count; $i++) {
3>   jq -r 'select(.text | tostring | contains(\"Dubrovnik\")) |
    [.text] | select(length > 0) | @csv' $files[$i].FullName
    >> "D:\twitter\results\Dubrovnik.csv"
4>   jq -r 'select(.text | tostring | contains(\"Rijeka\")) |
    [.text] | select(length > 0) | @csv' $files[$i].FullName
    >> "D:\twitter\results\Rijeka.csv"
5>   jq -r 'select(.text | tostring | contains(\"Plitvice\")) |
    [.text] | select(length > 0) | @csv' $files[$i].FullName
    >> "D:\twitter\results\Plitvice.csv"
6>   jq -r 'select(.text | tostring | contains(\"Plitvička\")) |
    [.text] | select(length > 0) | @csv' $files[$i].FullName
    >> "D:\twitter\results\Plitvice.csv"
7>  }
```

Line 1 initiates all the files contained in the specific directory. Line 2 begins a loop which contains 4 queries of jq. We invoke jq multiple times mainly due to the fact that we wanted to have mentions for different places saved in different files. The reader may also notice Plitvice shows up twice – this is because both the English and Croatian spelling of the name is included. Using the script shown above the program would go through the whole first day of month July. If hardware allows it and there is enough disk space it is possible to run the script for the whole month by omitting the last part of the directory path (\01\) on Line 1. All the shown scripts are completely and solely developed by the authors of this paper. We have not yet found any other solution to the discussed problems of automatic extraction.

Using Twitter as a rich source of information suitable for different research purposes is not a new or an unusual idea. If researchers choose to listen to public streams about upcoming political elections or major sport events all they have to do is set up a live streaming application using Twitter Track API and download data that matches the criteria. However, if the goal is to go back in time and collect tweets about past events couple of years ago, it is not possible to make a good thorough search in the Twitter archives the same way anyone can extract public streams due to the official limitations. There isn't public access available that allows researchers to dig through specific months in past years. If the goal is to compare user activity on a certain topic for specific months in the past years as of writing this paper it is not possible. The only way is to have researchers listening and monitoring the Twitter data streams in those past moments. Here our approach allows for full machine-automated extraction and search within the archived collections.

So, after the extraction was completed and the filtering algorithm was applied, we found that for July of 2016 there were *100 mentions of Rijeka, 110 for Plitvice and 248 for Dubrovnik.* For the same period next year (2017) *Rijeka had 50, Plitvice had 131 and Dubrovnik had 234.* It is important to note that those mentions come from original tweets, retweets or replies, that contain the filter words we chose. Furthermore, these are roughly 4% of all created tweets, so it is safe to assume the real number is around 20 to 25 times that. One step further into sentiment analysis and opinion mining of tweets would be to extract a mention with all the replies to that tweet. However, this can cause an exponential increase in processing time. For example, Ander Herrera, a soccer player from Manchester United tweeted #Dubrovnik [53], which caused 35 people to reply. In order to retrieve those replies one would need to extract the unique 'id' of his tweet and then search the archive for tweets that are replies to that `id`. However, some replies may come later, e.g. during the month of August, so in order to get better results one would need to have access to those tweets as well.

B. Using RapidMiner in conjunction with Twitter data

Text pre-processing, classification and the resulting visualization of the output are achieved using RapidMiner, a software solution that based on a Gartner research is among the best tools for data mining [54]. The software is a Java application that has a rich set of tools for pre-processing, classification, regression, clusterization, association and visual representation of results.

In order to reach our desired state of classification and to assess our approach we have prepared a significant amount of training data to teach the classifiers. In supervised machine learning it is crucial to have training data, so that the model can be validated. To help prepare the classifier and supply the supervised learning with initial data for the first classification stage two groups of text files are used – the first group contains neutral or objective statements, while the other has personal opinions on the studied subject. For the second classification stage again we prepared two groups of text files – this time containing positive and negative personal opinions.

The software of choice as already stated is RapidMiner and with it we build a model that will serve as a supervisor for the classifier (Fig. 6). Frist during the initial pre-processing of text data, vectors are built from words, using the class Process Documents from Files.

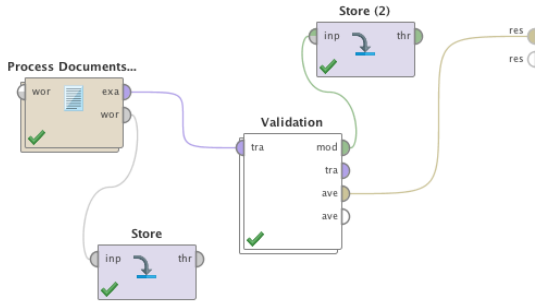

Fig. 6. Classification model, built in RapidMiner.

Text pre-processing can be broken down to the following micro processes (Fig. 7):

- replace tokens – for the replacement of abbreviations with their meaning;
- tokenization – for splitting the text of a document into a sequence of tokens;
- token filtering – based on their length;
- stemming – stems English words using the Porter stemming algorithm applying an iterative, rule-based replacement of word suffixes, the goal of which to reduce the length of the words until a minimum length is reached;
- stop words filtering – removes English stop words from a document;
- case transformation – transforms all characters in a document to lowercase.

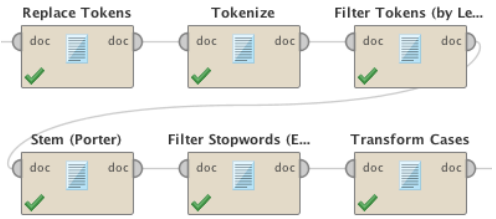

Fig. 7. Text pre-processing.

Transforming text messages into vectors requires looking at the occurrence rate of words. This task can be completed using a fairly common statistical algorithm: Term Frequency – Inverse Document Frequency (TF-IDF).

Evaluating the weight or significance of a word the following formulae are used:

$$tf(t,d) = \frac{1}{2} + \frac{\frac{1}{2}freq(t,d)}{\max\{freq(w,d):w \in d\}} \quad (3)$$

$$idf(t,D) = log\frac{|D|}{|\{d \in D: t \in d\}|} \quad (4)$$

$$TFIDF(t,d,D) = tf(t,d).idf(d,D) \quad (5)$$

where:

t – the term or a word;

d – the document in which the word is found;

D – the document corpus.

In general TF-IDF represents a statistical overview of the importance of words, in the context of current collection of documents or corpus. TF-IDF increases in value proportionally to the rate of which specific word occurs. It also takes into account that particular words will always occur more depending on the corpus.

To conduct the classifications, train the models and extract classification rules the Validation function is used, which allows the group segregation to occur based on an initially set algorithm. For the purpose of our research as mentioned earlier we will use two algorithms – SVM and Naïve Bayes.

To train the model the test data set is composed of sufficient enough number of elements – 600 in each group. When using the linear classifier SVM and the probability classifier Naïve Bayes the accuracy of the example data is tested as long as they match the data criteria for class precision and recall as shown in Table 1 and Table 2. The example dataset used in the classification training is verified beforehand and as a result the accuracy is very high.

TABLE I
OBJECTIVE / SUBJECTIVE CLASSIFICATION MODEL TRAINING DATA

	Classification with SVM		**Classification with Naïve Bayes**	
	Accuracy – 93,27%		Accuracy – 91,15%	
	Class precision	Recall	Class precision	Recall
Objective	95,43%	90,67	97,84%	86,32%
Subjective	91.83%	94,50%	88,03%	96,89%

TABLE II
POSITIVE / NEGATIVE CLASSIFICATION MODEL TRAINING DATA

	Classification with SVM		**Classification with Naïve Bayes**	
	Accuracy – 90,15%		Accuracy – 89,69%	
	Class precision	Recall	Class precision	Recall
Positive	84,23%	96,54%	96,67%	94,23%
Negative	93,78%	87,76%	84,73%	84,01%

To group the extracted from Twitter opinions about Croatian tourist destinations Rijeka, Plitvice, Dubrovnik we implemented the classification model inside RapidMiner as presented on Fig. 8.

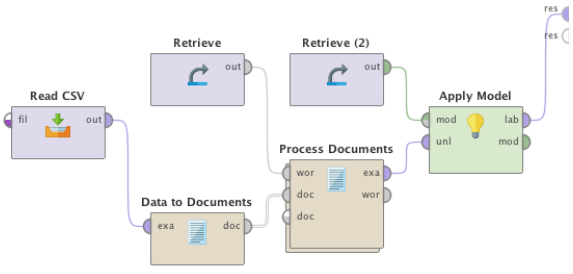

Fig. 8. Classification model implementation

The model we built reads the tweets from an Excel spreadsheet we created from the csv files that are generated using jq in the text pre-processing stage. Each tweet is transformed into a vector. After that the training classifier is applied to segregate the statements into Objective and Subjective. Then for each destination we split the opinions into positive and negative.

V. RESULTS

As a result of the extraction process for the tourist destinations Rijeka, Plitvice and Dubrovnik we identified for July of 2016 a total of 458 opinions (100 mentions for Rijeka, 110 for Plitvice and 248 for Dubrovnik), for July of 2017 another 415 opinions (50 for Rijeka, 131 for Plitvice and 234 for Dubrovnik) and for July of 2018 the last 532 opinions (47 for Rijeka, 184 for Plitvice and 301 for Dubrovnik) Descriptive statistics and cross tabulations may be retrieved using statistical analysis [48], but in this research machine learning methods are applied.

After processing the data using a linear classifier SVM for the three years the following results are obtained:

- for the year 2016, 27 objective statements (8 for Rijeka, 4 for Plitvice and 15 for Dubrovnik) have been identified, which are removed during the first stage of classification. The remaining 431 opinions are processed and there are 355 positive and 76 negative statements (27 for Rijeka, 8 for Plitvice and 41 for Dubrovnik).
- for the year 2017, 28 objective statements (3 for Rijeka, 10 for Plitvice and 15 for Dubrovnik) have been identified), which are removed during the first stage of classification. The remaining 387 opinions are processed, and there are 340 positive and 47 negative statements (11 for Rijeka, 13 for Plitvice and 23 for Dubrovnik).
- for the year 2018, 40 objective statements (7 for Rijeka, 11 for Plitvice and 22 for Dubrovnik) have been identified), which are removed during the first stage of classification. The remaining 492 opinions are processed, and there are 457 positive and 35 negative statements (10 for Rijeka, 13 for Plitvice and 12 for Dubrovnik).

After processing the data by applying the Naïve Bayes classifier for the three years the following results are obtained:

- for the year 2016, 34 objective statements (12 for Rijeka, 3 for Plitvice and 19 for Dubrovnik) are identified, while the remaining 424 opinions are split into 361 positive and 63 negative (25 for Rijeka, 5 for Plitvice and 33 for Dubrovnik).
- for the year 2017, 39 objective statements (7 for Rijeka, 13 for Plitvice and 19 for Dubrovnik) are identified, which are removed during the first stage of classification. The remaining 376 opinions are processed and as a result there are 324 positive and 52 negative statements (13 for Rijeka, 14 for Plitvice and 25 for Dubrovnik).
- for the year 2018, 39 objective statements (5 for Rijeka, 15 for Plitvice and 19 for Dubrovnik) are identified, which are removed during the first stage of classification. The remaining 485 opinions are processed and as a result there are 448 positive and 37 negative statements (7 for Rijeka, 16 for Plitvice and 14 for Dubrovnik).

In Table 3 a summary of the obtained results is made by destinations.

TABLE III
CLASSIFICATION RESULTS

	2016		2017		2018	
	SVM	**NB**	**SVM**	**NB**	**SVM**	**NB**
Rijeka						
% subjective statements	92%	88%	94%	86%	85%	87%
% positive statements	71%	72%	77%	70%	75%	79%
Plitvice						
% subjective statements	96%	97%	92%	90%	94%	92%
% positive statements	93%	95%	89%	88%	92%	91%
Dubrovnik						
% subjective statements	94%	92%	94%	92%	93%	94%
% positive statements	82%	86%	89%	88%	96%	95%

The performed test experiments show that the application of both SVM and Naïve Bayes algorithms gives similar results and they can be successfully used in textual classification. The application of different training algorithms [49] leads to finding the best one for the analyzed dataset.

We can conclude that based on the results for the three resorts, positive opinions predominate. It can be stated that the quality of services being provided in these holiday resorts is good. Based on the positive feedback we can expect in the next few years a rise in inquiries regarding tourist destinations in Croatia. This is due to the understanding that in general people are influenced by the opinions of others. We assume that satisfied customers lead to an increase in visitations in the following tourist seasons both by them and their friends, acquaintances and even people they don`t know personally.

When discussing the results of our research it is important to acknowledge that even though the proposed approach shows a good snapshot of users` attitude towards Croatian holiday resorts, there is a possibility of small inconsistencies due to irony or sarcasm in the tweets. These messages can be wrongly classified, but in a large set of comments their percentage would not be significant enough to skew the results.

The proposed approach can also be applied to a specific hotel or vacation resort, and the results would help the experts to analyze and evaluate customer reviews. Our algorithm can be developed and improved further by adding an analysis and detection of specific emotions in the text. To achieve this, a multi-class classification can be used, where opinions are classified exactly into one among many classes, representing particular emotions. As a subject of further research it is also viable to apply multi-label classification, to allow for attributing text into different classes at the same time. In general such tasks are broken down to a few binary and multiclass classifications or are resolved using specialized adaptive algorithms.

VI. CONCLUSION

The rapid development of social networking web sites and the opportunities they present for quick and easy sharing of information are important factors in the field of sentiment analysis and opinion mining. Everyone is constantly on the Internet, expressing their views and opinions making the web a huge source of subjective statements. Intelligent business analysis of customer impressions has been of great importance for corporations and researches in the field. Organizations rarely (if at all) share their know-how and so far, there has not been a firm confirmation in regard to a ubiquitous algorithm for text data mining. Thus, in our paper we propose and test an approach for computer analysis of user opinions, regarding holiday and vacation resorts. The resulting new knowledge, derived from web data can help hotel managers to better understand their customers and further improve their services, which of course can lead to economic growth in the sector. Data mining from social networks will inevitably become a vital part of the business strategy of any retailer of goods or services.

REFERENCES

[1] I. Kuyumdzhiev, "Controls Mitigating the Risk of Confidential Information Disclosure by Facebook: Essential Concern in Auditing Information Security Survey of Text Mining Techniques and Applications", *TEM Journal*, vol. 3, no. 2, pp. 113-119, 2014.

[2] R. Nacheva, "Social networks – a tool for providing a flexible learning process" in *Proceedings of Scientific Conference of Young Researchers 2013*, University publishing house "Science and economics", Varna: 2013, pp. 152-158.

[3] Olson, R., Neal, Z. "Navigating the massive world of reddit: Using backbone networks to map user interests in social media." *PeerJ Computer Science*, 1, 4, 2015.

[4] Serrano, M., Boguná, M., Vespignani, A. Extracting the multiscale backbone of complex weighted networks. "Proceedings of the national academy of sciences", 106(16), 6483-6488.

[5] Huang, J., Depari, G. "Paid Advertisement on Facebook: An Evaluation Using a Data Mining Approach." *Review of Integrative Business and Economics Research*, 8(4), 1, 2019.

[6] Twitter Usage Statistics, [Online]. Available: http://www.internetlivestats.com/ twitter-statistics/, [Accessed 24.12.2017].

[7] V. Gupta and G. S. Lehal. "A Survey of Text Mining Techniques and Applications", *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60-76, August 2009.

[8] U Fayyad et. al. "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, vol. 17, 1996, pp. 37-54, doi: 10.1609/aimag.v17i3.1230

[9] R. Feldman, J. Sanger. *The text mining handbook.* Advanced Approaches in Analyzing Unstructured Data, Cambridge: University Press, 2007.

[10] E. Kumar, *Natural Language Processing*, I. K., New Delhi: International Publishing House Pvt., 2011.

[11] L. Todoranova, „The creation and development of knowledge warehouses", *Izvestyia Journal of Union of scientists Varna*, pp. 156 - 161, 2015.

[12] A. Pena-Ayala, *Educational Data Mining*. Applications and Trends, Charm Heidelberg: Springer International Publishing, 2014.

[13] O. Etzioni, The World Wide Web: quagmire or gold mine?. *Communications of the ACM*, vol 11, 1996, pp. 65-68, doi: 10.1145/240455.240473

[14] R. Cooley et. al. "Web Mining: Information and Pattern Discovery on the World Wide Web", in: *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, 1997, pp. 558-567, doi: 10.1109/TAI.1997.632303

[15] F. Johnson and S. K. Gupta, "Web Content Mining Techniques: A Survey", *International Journal of Computer Applications*, vol. 47, no. 11, pp. 44-50, 2012. doi: 10.5120/7236-0266

[16] R. Kosala and H. Blockeel, "Web mining research", *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 1, pp. 1-15, 2000. doi: 10.1145/360402.360406

[17] D. Navadiya and R. Patel, "Web Content Mining Techniques-A Comprehensive Survey", *International Journal of Engineering Research & Technology*, vol. 1, no. 10, 2012.

[18] Z. Markov and D. Larosed, *Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage*. New Jersey: John Wiley & Sons.

[19] J. Materna, "Automated web page classification". in: *Proceedings of recent advances in Slavonic natural language processing*, Masaryk, Czech Republic, Masaryk: University Press, pp. 84-93, 2008.

[20] X. Qi and B. Davison, "Web page classification", *ACM Computing Surveys*, vol. 41, no. 2, pp. 1-31, 2009, doi: 10.1145/1459352.1459357

[21] S. Hariharan and R. Srinivasan, "A Comparison of Similarity Measures for Text Documents", *Journal of Information & Knowledge Management*, vol. 07, no. 01, pp. 1-8, 2008. doi: 10.1142/S0219649208001889

[22] S. Lakshmi, "Analysis of Similarity Measures for Text Clustering", *International Journal of Engineering & Science Research*, vol. 3, no. 8, pp. 4627-4639, 2013.

[23] E. D'Avanzo and G. Pilato, "Mining social network users opinions' to aid buyers' shopping decisions", *Computers in Human Behavior*, vol. 51, pp. 1284-1294, 2015, doi: 10.1016/j.chb.2014.11.081

[24] W. Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp. 1093-1113, doi: 10.1016/j.asej.2014.04.011

[25] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.

[26] 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. doi: 10.1016/j.asej.2014.04.011

[27] V. Patel, G. Prabhu and K. Bhowmick, "A Survey of Opinion Mining and Sentiment Analysis", *International Journal of Computer Applications*, vol. 131, no. 1, pp. 24-27, 2015, doi: 10.5120/ijca2015907218

[28] K. Dave, S. Lawrence and D. Pennock, "Mining the peanut gallery", in *Proceedings of the twelfth international conference on World Wide Web – WWW '03*, 2003. doi: 10.1145/775152.775226

[29] M. Sadegh et. al., "Opinion Mining and Sentiment Analysis: A Survey", *International Journal of Computers & Technology*. 2 (3). p. 171-178, 2012.

[30] S. Das and M. Chen, "Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web", *SSRN Electronic Journal*, 2001. doi: 10.2139/ssrn.276189

[31] R. Tong, "An Operational System for Detecting and Tracking Opinions in On-line Discussions", in *Working Notes of the SIGIR Workshop on Operational Text Classification*, pp. 1-6, 2001.

[32] D. Ankitkumar et. al., "A Survey on Sentiment Analysis and Opinion Mining", *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, 11, pp. 6633-6639, 2014.

[33] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", *Foundations and Trends® in Information Retrieval*, vol. 2, no. 12, pp. 1-135, 2008, doi: 10.1561/1500000011

[34] H. Rahmath, "Opinion Mining and Sentiment Analysis - Challenges and Applications", *International Journal of Application or Innovation in Engineering & Management*. 3 (5). pp. 401-403. 2014.

[35] P. Gonçalves and M. Araújo, "Comparing and Combining Sentiment Analysis Methods", Proceeding COSN '13 Proceedings of the first ACM conference on Online social networks, Boston, Massachusetts, USA, pp. 27-38, 2013, doi: 10.1145/2512938.2512951

[36] 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, doi: 10.1016/j.knosys.2015.06.015

[37] S. Sulova, "An Approach For Automatic Analysis Of Online Store Product And Services Reviews", *Izvestyia*, Varna University of Economics, issue 4, pp. 455-467, 2016.

[38] M. Hu and L. Bing, "Mining and summarizing customer reviews". in: *Proceedings of the 2004 ACM SIGKDD international conference on*

Knowledge discovery and data mining – KDD '04, pp. 168-177, 2004, doi: 10.1145/1014052.1014073

[39] S. Kim and E. Hovy, "Determining the sentiment of opinions". in: *Proceedings of the 20th international conference on Computational Linguistics – COLING '04*, 2004, doi: 10.3115/1220355.1220555

[40] R. Verma and Kiranjyoti, "Opinion Mining and Analysis of the Techniques for User Generated Content (UGC)", *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(5)/2015, p. 438-441.

[41] P. Singh and M. Husain, "Methodological study of opinion mining and sentiment analysis techniques". *International Journal on Soft Computing* (IJSC), 5(1)/2014, pp. 11-21, doi: 10.5121/ijsc.2014.5102.

[42] L. Lai and W. To, "Content analysis of social media: A grounded theory approach", *Journal of Electronic Commerce Research* 16(2)/2015, pp. 138-152.

[43] M. Scott, "Developing WordSmith", *International Journal of English Studies*, 8(1), 2008, pp. 95-106, doi: 10.6018/ijes.8.1.49111.

[44] H. Saggion and A. Funk, "Extracting Opinions and Facts for Business Intelligence", [Online]. Available: https://gate.ac.uk/sale/rnti-09/opinion-mining-09-v2.2.pdf, [Accessed 30.06.2019].

[45] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, S. Chernov, "Personalized Social Search Based on the User's Social Network", *CIKM '09* Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 1227-1236, doi: 10.1145/1645953.1646109.

[46] Z. Cheng, J. Caverlee, K. Lee, You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users, CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 759-768, doi: 10.1145/1871437.1871535

[47] H. Gao and H. Liu "Data Analysis on Location-Based Social Networks". In: *Chin A., Zhang D. (eds) Mobile Social Networking. Computational Social Sciences*. Springer, New York, NY, 2014, doi: https://doi.org/10.1007/978-1-4614-8579-7_8.

[48] *91 Free Twitter Tools and Apps to Fit Any Need*, [Online]. Available: www.slideshare.net/Bufferapp/91-free-twitter-tools-and-apps-to-fit-any-need, [Accessed 30.06.2019].

[49] *Archive Team: The Twitter Stream Grab* [Online]. Available: https://archive.org/details/twitterstream&tab=about, [Accessed 30.12.2017].

[50] T. Wilson, et. all, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis" in: *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347-354, 2005, doi: 10.3115/1220575.1220619

[51] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", in: *Proceeding ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004, doi: 10.3115/1218955.1218990

[52] I. Kuyumdzhiev, "Comparing Backup and Restore Efficiency in MySql, MS SQL Server and Mongodb", in *19 International Multidisciplinary Scientific Geoconference SGEM 2019*, Albena, Bulgaria, vol. 19, issue. 2.1, 2019, pp. 167-174.

[53] *Ander Herrera on Twitter*, [Online]. Available: https://twitter.com/AnderHerrera/status/747521324650795008/photo/1, [Accessed 30.12.2017].

[54] L. Kart, et. all, Magic Quadrant for Advanced Analytics Platforms, [Online]. Available: https://rapidminer.com/resource/gartner-data-science-platforms-magic-quadrant/, [Accessed 28.12.2017].

[55] A. Tarasyev, J. Vasilev, V. Turygina, "Statistical analysis and forecasting of extraction and use of natural resources" in AIP Conference Proceedings, 2018, pp. 050011-1–050011-4. doi: 10.1063/1.5079109.

[56] A. Shichkin, A. Buevich, A. Sergeev, E. Baglaeva, I. Subbotina, J. Vasilev, M. Kehayova-Stoycheva, "Training algorithms for artificial neural network in predicting of the content of chemical elements in the upper soil layer" in *AIP Conference Proceedings*, 2018, pp. 060004-1–060004-5. doi: 10.1063/1.5082119.

Snezhana Sulova was born on August 21, 1973 in Nova Zagora, Bulgaria. She received a Ph.D. degree in 2005. In 2006 she joined the Computer Science department of University of Economics – Varna as an assistant. During the period 2008–2012 she was a Chief assistant professor, and since 2012 she is an Associate Professor In the Computer Science department of University of Economics.
Her research interests are in the field of Data Mining, Text Mining, Machine Learning, Artificial Intelligence, E-commerce.

Boris Bankov was born on March 18, 1990. He received B.Sc., M.Sc and PH.D. degrees in IT and Computer Science from University of Economics – Varna in 2013, 2014 and 2018 respectively. In 2014 he joined the Computer Science department of University of Economics – Varna as an assistant and in 2019 he advanced to the position of Chief assistant. His research interests are in the field of unstructured data processing, clusterization, neural networks, deep learning, gamification.