Brano Markić
University of Mostar
Faculty of Economics
Matice Hrvatske b.b.,
88000 Mostar,
Bosnia and Herzegovina
brano.markic@sve-mo.ba
Phone:+38736355100

Sanja Bijakšić
University of Mostar
Faculty of Economics
Matice Hrvatske b.b.,
88000 Mostar,
Bosnia and Herzegovina
sanja.bijaksic@sve-mo.ba
Phone:+38736355100

Arnela Bevanda
University of Mostar
Faculty of Economics
Matice Hrvatske b.b.,
88000 Mostar,
Bosnia and Herzegovina
arnela.budimir@sve-mo.ba
Phone:+38736355100

UDK: 004.7:339.138
Preliminary communication

Received: November 11, 2015
Accepted for publishing: March 14, 2016

SENTIMENT ANALYSIS OF SOCIAL NETWORKS AS A CHALLENGE TO THE DIGITAL MARKETING

Abstract

Huge amounts of data, in the form of messages on social networks, represent a challange for digital marketing and marketing analytics when meeting the requirements, needs and customer satisfaction with services or products. Marketing strives to be a part of the overall culture based on the data and to define marketing strategies that respond to consumers and thus to provide economic benefits for the company. Therefore, the focus of marketing analysis is on the data recorded at the social networks. This paper shows one possible integration of information technology and data mining tools, with the goal of visualizing the attitudes and opinions on the social networks in the form of a word cloud, which can then further be used to create marketing strategies and improve customer relations and customer service.

Keywords: Text mining, word cloud, marketing analytics, R language, data mining

1. Introduction

Information about products, services, customers, suppliers and transactions are written and stored in the form of a database within each company. These are large quantities of detailed information that cannot be directly used for analytical purposes; therefore, they are aggregated into a dimensional model of data warehouse. These data can be reached and analyzed by various data mining algorithms in order to extract regularities or laws hidden in a dimensional model. Information layout is pre-determined, i.e. it is designed and structured to be automatically analyzed by data mining algorithms. However, marketing nowadays, in the analysis of customer behaviour and identifying their needs and requirements cannot be satisfied with such formatted data in the database or data warehouse. The best example are social networks as a virtual space for the exchange of opinions, feelings, requirements, motives, ideas, views about companies, products, services, events, and destinations. The data on social networks are in a text form, and its layout is not previously determined (it is not formatted as a database or data warehouse). Business requires real-time information. The main hypothesis in the paper is that the sentiment analysis which results as a word cloud is simple but very informative and usable for its users (marketing experts). Marketing experts are faced daily with the collection and analysis of data about customers and products in order to achieve competitive advantages. Sentiment analysis or opinion mining is related to an application of natural language processing, computational linguistics and text analytics. Therefore, digital marketing is being intensively developed through data analysis marketing, information technology and software tools, which tends to use data on products and services that are registered on social networks in order to form marketing ace and marketing strategy. The infrastructure of a global network has enabled development of various types of electronic marketing including viral marketing, affiliate marketing, real-time marketing, one-to-one marketing, e-mail marketing, referral marketing, permission marketing and frequency marketing. The user (customer) forms a website, a blog and becomes a part of social networks: (Twitter, Facebook, YouTube, Instagram, LinkedIn and many others). By using these forms of web technologies, one 'leaves' useful data for marketing analysts about events, personal impressions on products, services, quality or product defects, consumers' needs and desires.

The purpose of the opinion analysis on social network users is to generate usable information for designing and implementing marketing strategies of companies.

1.1 Sentiment analysis

Companies achieve success or failure in the market due to customers. The purpose of each company is therefore to meet the needs and desires of their customers, their preferences and purchasing habits. The experience of successful companies has shown that only those who know their customers can satisfy their needs and so create loyal customers (Bijakšić et al., 2014).

Surveys of users and cusomers is a qualitative analysis that is directly related to psychology due to its focusing on emotions (feelings) that are formed when using products or services. Emotions are directly associated with the conscious, but also unconscious part of the human system. Sentiment analysis has been present for a long time. It has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole (Bing 2010). However, the research on sentiments and opinions appeared earlier (Morinaga et al., 2002; Tong, 2001; Turney, 2002; Wiebe, 2000).

The instruments for data collection on opinions and attitudes of customers about a product are mostly questionnaires, interviews or direct comments (customer comment cards, surveys, interviews). In addition to these traditional instruments for collecting data about customers and their opinions, marketing today is also focused on social networks. Messages on social networks are systematically reviewed and opinions of customers are recorded. Some examples of messages that express opinions of customers about a tourist destination, a product or a company are:

'Mostar is a beautiful, clean and promising city.'

'Bread of X manufacturer is not tasty.'

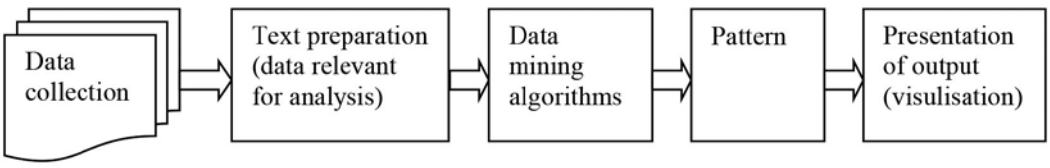
'Company Y has a high-quality service.'

Detecting opinions of customers has tremendous practical value for the company. Therefore, senti-

ment analysis is a challenge for marketing analysis because it can directly contribute to customer's satisfaction and thus the company's success on the market. It allows the company's proactive actions and it can direct operational activities of companies where properties and quality of products are adapted to the requirements of the customers. The application of sentiment analysis is not simple. It is always the integration of marketing, IT, mathematical and linguistic knowledge, which results in information on customers' opinions, their satisfaction and needs. Bing defines sentiment analysis as a process in which text data finds attitudes, feelings and emotions about an object or issue (Bing, 2010). Therefore, sentiment analysis cannot be separated from mathematical and algorithmic approach that finds hidden content and information from data, i.e data mining.

According to many marketing experts and analysts, special attention is drawn to extracting opinions and attitudes regarding text messages. In terms of methodology, sentiment analysis is a series of sequential steps. These steps follow one another but it is not possible to implement such a linear sequence in practical terms. Sentiment analysis begins with data collection. The data is in the form of a text on social networks. After collecting the text data it is necessary to prepare them for the analysis and thus detect the opinions and attitudes of social network users.

Figure 1 Steps in the analysis of attitudes and opinions (sentiment analysis) on social networks



Source: Authors

Analysis of views and opinions is seen as the process where text data creates knowledge using natural language processing, text mining algorithms and machine learning.

Figure 1 is a conceptualization of the overall concept of sentiment analysis, split into five stages. It begins with collecting the data on social networks and ends with the presentation of data (usually in the visual form).

Data collection is followed by their cleaning, prior to analysis by a data mining algorithm. Preparation of the text is the process of eliminating stop words (e.g. - ; , : ?) or words that are not relevant to the analysis.

In analyzing the data, only the messages that contain a subjective opinion on the event, process, person, destination, i.e. object of analysis, will be retained in the data set.

The purpose is to translate the documents into simpler form so that they are suitable for parsing. Sentiment analysis can concentrate on the message, the word, or part of the message.

In the analysis, unigram, n-gram, lemme, negations, words that express an opinion can be used. Unigram shows each element as a vector of features that contains the frequency of its occurence in the document. N-grams are similar as containing several words in a row and they affect a wider context. Lemmas (lema) are used as synonyms, for example good-better-the best. Data mining algorithms retrieve data and perform their processing. This can be, for example, classification into particular groups, where Naive Bayes algorithm or maximum entropy is applicable. This research analyzes the results presented in the form of a word cloud. It is a method of data visualisation when high frequency words are written in the biggest font and reflect the most common terms that appear in the text. By linking these words it is possible to detect opinions about a particular product or destination. Such analysis can then be used for the application of appropriate measures of sales promotion, advertising and other marketing instruments to improve the image or retain the existing reputation, and to attract new customers.

2. The role of sentiment analysis in marketing

Marketing constantly tends to analyse customers, products, prices, suppliers, products or services in order to satisfy needs and desires of customers.

It is possible to fulfil such a task only by permanent (continuous) study of attitudes and opinions, measuring satisfaction, analysis of the effectiveness, promotions, and the like.

There are two basic forms of research: qualitative and quantitative.

Qualitative market research for example, tends to know the perception of the product or services by customers. Sometimes, the food product can have the highest quality components, the latest technology standards, but the taste of the product is still not acceptable to customers. The customer makes the assessment of satisfaction and its quality only on the basis of parameters which are the result of qualitative research. Qualitative research is always trying to figure out how the customer feels and what (s)he thinks. Quantitative analysis uses a large number of data, relies on large samples and tends to figure out tendencies to give further conclusions and recommendations (Markić, 2014).

The relations between quantitative and qualitative analyses are not simple, therefore, discussions on credibility and applicability of either approach are conducted. The market research indicates that both of these have their applicable areas. There is an increasing number of researches which integrate both qualitative and quantitative analysis. Sentiment analysis has both of these components. The data that are analyzed are in the text format, but conclusions about their meanings, patterns and regularities are made by applying the quantitative analysis techniques. Quantitative techniques within sentiment analysis allow 'deep' and credible insights regarding opinions and feelings of the service or product user.

Application of data mining or machine learning provides strength and scientific objectivity and credibility to qualitative analysis. Users of this analysis ger an access to real-time opinions, attitudes and feelings of service users. Therefore, sentiment analysis has a special value for analysing and getting insights into the image of tourist destinations.

A destination image is the totality of perceptions, attitudes, knowledge, experience, expectations, desires and emotions of people about a particular tourist destination.

The creation and promotion of the "favorable" image of a destination among potential tourists has a special significance.

Based on the significance of the image of tourist destinations, destination marketers should know how to manage the destination image in terms of its creation, enhancement or modification, and in this sense it is necessary to learn about the processes of formation, measuring and analyzing the image of a tourist destination.

Different information and its sources on social networks have a big importance to a communication process in the context of the image of the tourist destination. It is becoming increasingly important to intensify the competitive relationship between tourist destinations, such communication processes and promotional efforts, in the context of creating the image of a tourist destination. These facts suggest differentiating the marketing of tourism destination from other activities.

There are many determinants of the image of the tourist destination, and among them we choose Gartner's key theses:

1. people who live in different geographical areas will have different images of the same destinations
2. if a potential tourist with their permanent residence is farther from a tourist destination, his image will be less clear
3. image is changing very slowly, so that only a serious change in the image of a tourist destination can change the perception of a specific destination
4. if the designated tourist destination is smaller, it is more likely to dominate the image of the political entity in which it is located
5. image is formed and continuously modified in different ways.

The data for the analysis of customers and service users are located in a natural, cyber environment. The data, therefore, have a special quality, as researchers have no effect on forming respondents' opinions. In addition, data collecting through a questionnaire always leaves the impression on respondents that the answers will be analyzed, and as a rule, this affects the objectivity of their answers.

The quality of data in the cyber environment, on social networks, is better compared to data collected in a questionnaire. The data is structured in a questionnaire because the respondents answer to something that is already predetermined and messages are completely spontaneous, natural and reflect their real opinion. Of course, the analysis of data collected on the basis of the messages is more complex.

3. Research methodology

Our researsch is concetrated on the analysis of Twitter users' opinions and attitudes on Mostar as a destination. The aim is to build the software application that can access the messages on a social networking site Twitter and to visualize keywords about Mostar as a destinaton in a graphical form. It is necessary to explore the possibilities of building software to access messages, watching them as documents and then extract the views and opinions of users of the tourist destination. The basic hypothesis is that it is possible to build an application that will ensure a marketing expert opinions about a tourist destination and its image.

Based on theoretical knowledge in building such software applications its results will be tested on the example of Mostar as a tourist destination.

The image of a tourist destination for more than 30 years has been the focus of scientific research. In the literature we find many definitions of tourist destinations which agree that the image of a destination is "a set of beliefs, ideas and impressions that people have about the place or the destination" (Sudar, 1991).

3.1 R language in application of sentiment analysis and a word cloud about tourist destination

R language started as a project by Ross Ihaka and Robert Gentleman at the Department of Statistics, University of Auckland, New Zealand, during the 1990s. The primary goal of this project was creating a statistical environment in their teaching lab. The international "R-core" team of some 15 people with access to the common CVS archive has been active since 1997. R is an interpreted computer programming language where most user-visible functions are written in R itself. It is also possible to interface procedures written in other languages (C, C+, or FORTRAN). R language is used for data manipulation, statistics, and graphics. R is made up of:

- operators (+ - <- * %*% ...) for calculations
- a collection of functions for making quality graphics and sets of functions (packages).

R is simple and also suitable for data analysis. For example, a linear regression model for relationship between the weight and height of students (data are stored in a data set student) includes very intuitive and simple statements:

```
erste.lm <- lm (weight ~ height, data = student) #
lm means linear-model
```

R can be used as a calculator for complex mathematical functions which is illustrated by another simple example[1]:

```
> exp(2)
[1] 7.389056# calculate and show
> (log (2)-log (4)) / log (2 ^4)
[1] -0.25
```

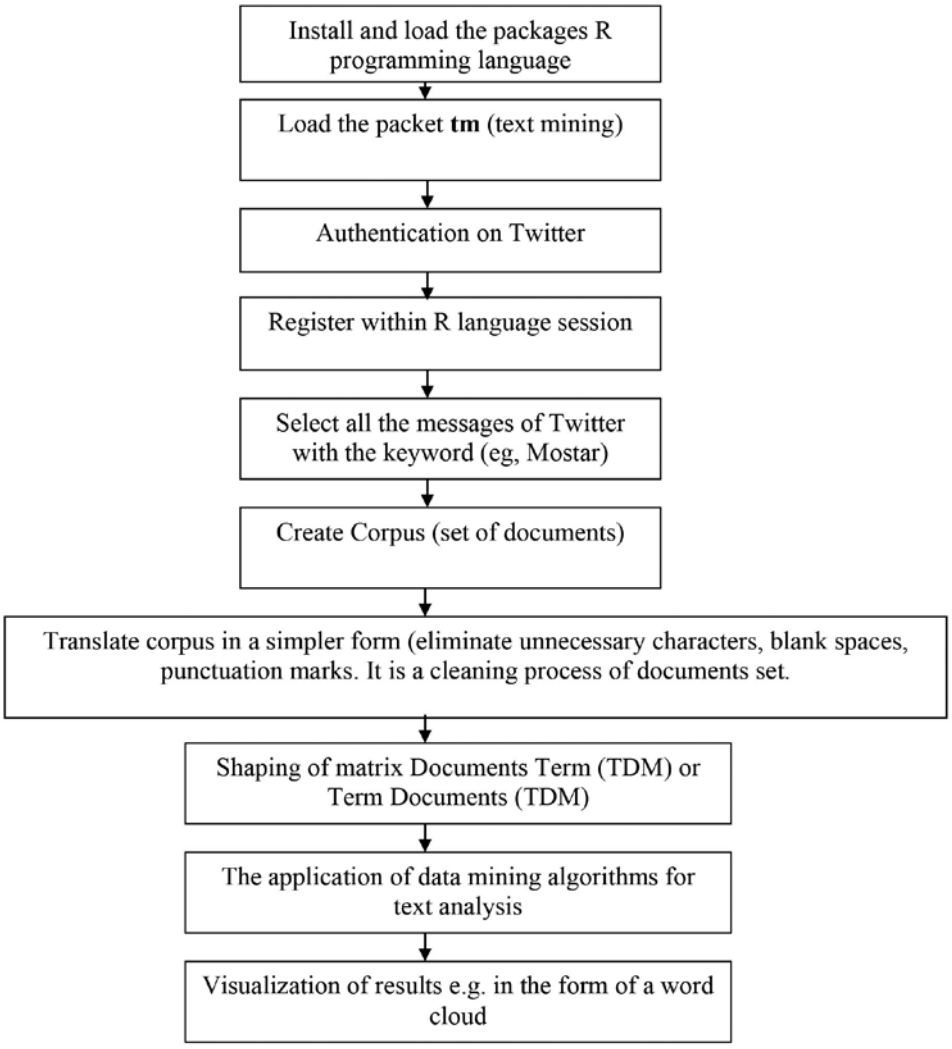
R is a programming language. It has the logical structure of programming: loops, branching, subroutines. It creates functions, classes, graphics and is associated with other programming languages such as C + +, Java, Perl, etc. Definitely, R language is not just a calculator. It has many built-in statistical and graphing functions and connects to databases and spreadsheets.

Particularly, data frames are of great importance. A data frame in R language is a collection of related vectors. The columns of the data frame can have different data types (mode) and most of the time, when data is loaded, it will be organized in a data frame[2].

Steps and commands in the language R will be shown, which are selected by the message in which Mostar is the keyword, and then the words that have the highest frequency in the messages will be displayed.

Input data are on the social network Twitter. The number of messages related to a particular destination is unknown in advance and that number is variable over time. Access to data (messages) stored on the server of Twitter allows proper function of the R language.

Figure 2 Steps of the algorithm for detecting the views and opinions on the social network by using the R programming language



Source: Authors

This function is key for selecting messages about a tourist destination and then for their analysis. The assumption is that a larger number of messages are related to attractiveness of the destination. If we cannot find any messages about the particular destination, then it is reasonable to conclude that the destination is not attractive. Access to messages is possible by using appropriate commands and functions of R programming language.

The next diagram shows the steps of the algorithm to analyze the attitudes and opinions expressed on social networks by using the R programming language.

Matrix is also important in text mining. In R language the matrix is a rectangular table with data of the same type.

3.2. Creating of word cloud for tourist destinations

Research methodology of opinions on social networks in terms of information technology has to be transformed into an algorithm. An algorithm is a set of unambiguous steps that lead to the solution of problems. It most commonly appears in the form of a block diagram (flowcharts) and reflects the natural way of thinking about the problem and its solution. This algorithm is a sequence program structure ie. a series of steps. Every step of the algorithm refers to the calling of the appropriate packages and functions in the programming language R.

Choosing the word cloud is a result of several reasons because of its several useful applications in marketing. Word cloud helps you to understand your customers, how they view you and what the most common impressions of your brand really are. Also, word cloud as a form of qualitative analysis in digital marketing enables you to identify a product, event, tourist destination or a brand whose online presence is popular with the customers that you want to reach. Ease of presentation and the possibility of a graphic display, a quick understanding of the results of qualitative marketing analysis are important reasons for widespread deployment of the word cloud in digital marketing. The flow chart of algorithms used for creating a word cloud and detecting the most frequent words connected with the choosen term (in our example the choosen term is tourist destination) is shown in Figure 2.

The flowchart (Figure 2) shows that the first step is loading packages for access to the social network Twitter and user authentication. Just write the following two commands:

```
> library(twitteR)
> library(ROAuth)
```

The process of authentication is necessary for all transactions on Twitter and it provides the full functionality of the application for sentiment analysis about the destination. The first step of authentication is login to the URL https://twitter.com/apps/new, then fill in the necessary data. It follows the entry of new instructions, but now in the development environment of R language:

```
> cred <- OAuthFactory$new(consumerKey = Vaš Key,
+ consumerSecret = Vaš SECRET,
+ requestURL = "https://api.twitter.com/oauth/request_token",
+ accessURL = "https://api.twitter.com/oauth/access_token",
+ authURL = "https://api.twitter.com/oauth/authorize")
> library(RCurl)
> download.file(url="http://curl.haxx.se/ca/cacert.pem",+ destfile="cacert.pem")
> cred$handshake(cainfo="cacert.pem")
```

To enable the connection, please direct your web browser to:

https://api.twitter.com/oauth/authorize?oauth_token=yF_07QAAAAAAhW7PAAABT5JyCPc

Then enter the PIN as it is written on the top of the URL.

When complete, record the PIN given to you and provide it here:

Allowing access to Twitter messages is possible only if the session is registered within the R language using the function:

```
> registerTwitterOAuth(cred)
[1] TRUE
```

For use in new applications it is useful to store the process of authentication in a separate file and reuse, which is possible by using the command:

```
> save(cred, file="twitter authentication.Rdata")
```

The next step is to extract all messages of Twitter with the keyword *Mostar* as a destination (searchTwitter function ()).

```
> options(RCurlOptions = list(verbose = FALSE, capath = system.file("CurlSSL", "cacert.pem", package = "RCurl"), ssl.verifypeer = FALSE))
>dataM<-searchTwitter("Mostar", n=1000)
```

Warning *message*:

In doRppAPICall("search/tweets", n, params = params, retryOnRateLimit = retryOnRateLimit, :

1000 tweets were requested but the API can only return 546.

The total number of messages that contain the name of the destination Mostar is 546 so it is in the R session visible warning messages (in the previous statement 1000 messages are defined). Twitter search and collection of data about the attitudes on the social network is in the form of textual data. These data are not structured i.e. their shape is not known in advance. To work with textual data the R language uses special package ***plyr*** which contains the ***laply()*** function to transform the messages in the text:

```
>library(plyr)
>twText<-laply(dataM,function(t) t$getText())
> head(twText,4)
[1] "Divisions threaten #Bosnia's prospects for #EU membership, #BBCNews reports: http://t.co/NIn1N4xSBR #BiH #Croatia #Sarajevo #Mostar #jmbg"
[2] "Future brother in law :-P #Mostar #summer #friends #nikes #2013 #Monday http://t.co/imneL9j1FK"
[3] "Plemići sve jači: Savić, Popović i Brković novi igrači Zrinjskog #Zrinjski #Mostar #Hercegovina http://t.co/1OV4Hi9L8W @radiosarajevo"
[4] "#mostar #bridge #bosna @ Stari Grad | Old Town http://t.co/ytY5m0wbPP"
```

The next step is to load the package tm (abbreviation of English words text mining). Description of the package is provided in the Journal of Statistical Software. The main structure for handling documents in the package ***tm*** makes the so-called Corpus and it represents a set of documents. You must enter the following sequence of commands:[3]

```
> library(tm)
>mostarCorpus<-
Corpus(VectorSource(twText2))
```

After being "caught" a collection of documents and all Twitter messages that contain the keyword Mostar (mostar Corpus) we need to modify the document so that it will eliminate the blank spaces in the documents (messages), signs of punctuation and the like. The logic is to eliminate all signs and words that do not carry information relevant to the semantics of the text.

The aim is to translate the documents into a simpler form so that they are suitable for parsing. It is sufficient in the loop for ... add commands that from the documents extract characters or words that will be replaced by spaces. It displays the following loop:

```
for(i in seq(docs))
 {
 mostarCorpus [[i]] <- gsub("/", " ", mostarCorpus [[i]])
  mostarCorpus [[i]] <- gsub("#", " ",  mostarCorpus [[i]])
  mostarCorpus [[i]]  <- gsub("\\|", " ", mostarCorpus [[i]])
  mostarCorpus [[i]]  <- gsub("kao", " ", mostarCorpus [[i]])
  mostarCorpus [[i]]  <- gsub("a", " ",  mostarCorpus [[i]])
  mostarCorpus [[i]]  <- gsub("je", " ", mostarCorpus [[i]])
 }
```

To delete the empty spaces it is sufficient to call the function ***stripWhitespace***, for punctuation function ***removePunctuation*** or for digits the function ***removeNumbers*** , which displays the following series of statements[4]:

```
mostarCorpus<- tm_map(mostarCorpus, stripWhitespace)
mostarCorpus<- tm_map(mostarCorpus, removePunctuation)
mostarCorpus<- tm_map(mostarCorpus, removeNumbers)
```

The next step is to create a matrix for learning which includes frequency of terms.

```
>tw.tdm<- TermDocumentMatrix(twCorpus, control = list(minWordLength = 5))
>inspect(tw.tdm)
```

A term-document matrix (108 terms, 1 document)

The number of terms in the given example is 108.

A document-term matrix[5] or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of messages from Twitter about Mostar as a tourist destination. In a document-term matrix, rows correspond to documents in the collection and columns correspond to Terms.

The terms (terms) are chosen based on the frequency of the document (Document Frequency-DF). Attributes that exceed a certain threshold will form a list of index terms. To display the results of processing it is necessary to load the package word cloud in the R session. It follows the formation of the matrix Term Documents and sorting rows in descending order. The rows of the matrix ***mostar.tdm*** are Terms and the columns are the documents (messages).>library(wordcloud)

```
>mostar.tdm <-
TermDocumentMatrix(twCorpus)

>m.m <- as.matrix(mostar.tdm)

>m.<- sort(rowSums(m.m),decreasing=TRUE)

>m.d<- data.frame(word =
names(m.v),freq=m.v)
```

Finally, only two commands of R language are sufficient to display in color and in the form of a word cloud terms that reflect the views and opinions about Mostar as a destination on Twitter.

```
>colWC <- brewer.pal(8,"Dark2")

>wordcloud(m.d$word,m.
d$freq,scale=c(8,.2),min.freq=4,max.words=Inf,
random.order=

FALSE, rot.per=.15, colors=colWC)
```

Figure 3 Word cloud analyses opinions on Mostar as a destination

Source: Authors

In the figure above the words that have the highest frequency in messages containing Mostar as a keyword are displayed. Since the word Mostar appears the most (in each message), that word has the largest font in the word cloud. Analyzing the word cloud it can be concluded that the words associated with Mostar as a destination are the following: old, bridge, beautiful, Bosnia, Herzegovina, photo, which suggests that Mostar in Herzegovina is a nice town where the Old Bridge is a historical element, where visitors like to take pictures. Of course, opinions of social network users depend on the point in time when the analysis is performed.

Figure 4 Sentiment analysis on Mostar as a destination

Source: Authors

For example, the analysis of attitudes and opinions on Mostar expressed on Twitter since May 2013, in the form of a word cloud, indicates the social moment regarding the political scene which was dominated by protests and the personal identification number of citizens, as shown in Figure 4.

4. Conclusion

In the scientific community there is an interest in sentiment analysis, but also in the main area of both marketing business function and marketing process that is increasing on a daily basis. This is a new research area in which marketing experts, IT specialists, marketing analysts, mathematicians and psychologists are confronted and where they cooperate at the same time. The data in the text form is specific in terms of content that is directly related to a specific time and space. In-depth analysis of the text

mining is based on the analysis of quantitative techniques where the text is adapted in order to apply quantitative methods of machine learning methods. There can be some methods that allow classification such as Bayesian analysis, cluster analysis or the nearest neighbour method. Quantitative techniques within sentiment analysis allow 'deep' and credible insights regarding opinions and feelings of the service or product users. Sentiment analysis which results in a word cloud is shown as simple, but very informative and applicable. The area of application is marketing because it allows detection of preferences, attitudes and opinions on social networks regarding products, services, companies and destinations. The special value of this analysis is the visualization in the form of word clouds where high frequency words are more centrally positioned and are displayed in font size proportionate to the number of times the word was mentioned in the text.

This study presents steps of sentiment analysis and all the essential functions of the R programming language which provide sentiment analysis on the example of Mostar as a destination. Its contribution to the conversion of large amount of textual data in real time is very useful for marketing analysis.

References

1. Bijakšić, S., Bevanda, A., Markić, B. (2014). Marketing i metrika, marketinški splet, podaci i mjerila. HKD Napredak - Glavna Podružnica Mostar.
2. Bijakšić S., Markić B., Bevanda A. (2013), "Text mininig and analysis of attitudes and opinions on social networks", Conference Proceedings of the International Conference: Tourism today for tomorrow, Šibenik.
3. Bing, L. (2010), Sentiment Analysis and Subjectivity, Invited Chapter for the Handbook of Natural Language Processing, Second Edition, Available at: http://gnode1.mib.man.ac.uk/tutorials/NLP-handbook-sentiment-analysis.pdf (Accessed on: September 12, 2015)
4. Davis, J., Shannon, O'F. (2012), "Assessing the Accuracy of Automated Twitter Sentiment Coding", Academy of Marketing Studies Journal, Vol. 16, No. S1, pp. 35-50.
5. Markić, B. (2014). Sustavi potpore odlučivanju, podaci, modeli i algoritmi. HKD Napredak - Glavna podružnica Mostar.
6. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T. (2002), "Mining product reputations on the web", in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002).
7. Sudar, J., Keller, G., (1991). Promocija. Zagreb: Informator.
8. Tong, R. M. (2001), "An operational system for detecting and tracking opinions in on-line discussion", in Proceedings of SIGIR Workshop on Operational Text Classification.
9. Turney, Peter D. (2002), "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", in Proceedings of Annual Meeting of the Association for Computational Linguistics.

Endnotes

1 Results of calculations can be stored in variables (in terminology of object oriented programming language in objects) using the assignment operators:

An arrow (<-) formed by a 'smaller than' character and a hyphen without a space!

The equal character (=).

```
> a <- 5                # can define variables using "<-" operator to set values.
>b=6                    # using "=" operator to set values
> 1 <- 4
> a * b * c
[1] 30
>A * B * C              # variable names are case sensitive
Error: Object "X" not found
> This.Year <- 2015# variable names can include period
> This.Year
[1] 2015
> 5 %/% 3                          # integer division
[1] 1
> 5 %% 3                           # modulo division
[1] 2
> exp(log(9)) # log() is the natural logarithm
[1] 9
> sin(pi/2)                        # sine
[1] 1
> cos(pi/2) # cosine of pi/2 is zero. Note: R does not answer with zero!
[1] 6.123234e-17
> factorial(5)                     # 4 factorial
[1]  120
Most sources define the binomial coefficient (n, k) as
> choose(6,3)                      # 6 choose 3
[1] 20
```

2 The function ***data.frame*** used to generate this data set.Number of columns is obtained by function length ()and dimension by function ***dim ()***. One example of data frame is

```
a<-c(seq(1:4))
> b<-c(58,65,77,84)
> c<-c(TRUE,FALSE,FALSE,TRUE)
> data.f<-data.frame(a,b,c)
> data.f<-data.frame(a,b,c)
> names(data.f)<-c("ID","weight","overweight")
> data.f
ID weight overweight
1  2     58       TRUE
2  2     65      FALSE
3  3     77      FALSE
4  4     84       TRUE
```

3 To understand the statements of R language it is helpful to visit websites that give detailed explanations of the use of the package and function in the R language. For package tm useful site is: http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf.

4 To convert all uppercase letters to lowercase letters it is enough to call the function tolower (twCorpus = tm_map (myCorpus, tolower) and determination of root words in the English language allows function stemDocument:

twCorpus = tm_map(myCorpus, stemDocument).

A corpus of documents contains all the words that appear in the documents. The term-document matrix then is a two-dimensional matrix whose rows are the terms and columns are the documents, so each entry (i, j) represents the frequency of term i in document j.

In information retrieval inverse document frequency (tf–idf) is used. For each entry in the matrix, the term frequency measures the number of times that term i appears in document j. The inverse document frequency measures the number of documents in the corpus which contain term i. If tf is the number of times term t appears in a document, and t is the total number of terms in the documents, then the term frequency of term t is:

tf(t)=tf/t

In tf(t) indicator all terms are considered equally important. However, it is known that certain terms may appear a lot of times but have little importance. Thus the idf (Inverse Document Frequency) indicator is used:

idf(t) = log_e(total number of documents (messages) / number of documents (messages) with term t in it).

A corpus of documents contains all the words that appear in the documents. The term-document matrix then is a two-dimensional matrix whose rows are the terms and columns are the documents, so each entry (i, j) represents the frequency of term i in document j.

In information retrieval inverse document frequency (tf–idf) is used. For each entry in the matrix, the term frequency measures the number of times that term i appears in document j. The inverse document frequency measures the number of documents in the corpus which contain term i. If tf is the number of times term t appears in a document, and t is the total number of terms in the documents, then the term frequency of term t is:

tf(t)=tf/t

In tf(t) indicator all terms are considered equally important. However, it is known that certain terms may appear a lot of times but have little importance. Thus the idf (Inverse Document Frequency) indicator is used:

idf(t) = log_e(total number of documents (messages) / number of documents (messages) with term t in it).

The tf-idf score is the product of these two metrics (tf*idf). This tf-idf score is ranking highly the documents which contain with high frequency the terms in the search query.

5 A corpus of documents contains all the words that appear in the documents. The term-document matrix then is a two-dimensional matrix whose rows are the terms and columns are the documents, so each entry (i, j) represents the frequency of term i in document j.

In information retrieval is using inverse document frequency (tf–idf). For each entry in the matrix, the term frequency measures the number of times that term i appears in document j. The inverse document frequency measures the number of documents in the corpus which contain term i. If tf is the number of times term t appears in a document and t is total number of terms in the documents then the term frequency of term t is:

tf(t)=tf/t

Intf(t) indicator all terms are considered equally important. However it is known that certain terms may appear a lot of times but have little importance. Thus is using the idf (Inverse Document Frequency) indicator:

idf(t) = log_e(total number of documents (messages) / number of documents (messages) with term t in it).

The tf-idf score is the product of these two metrics (tf*idf). This tf-idf score is ranking highly documents which contain with high frequency the terms in the search query.

UDK: 004.7:339.138 / **Preliminary communication**

Brano Markić
Sanja Bijakšić
Arnela Bevanda

Sentiment analiza društvenih mreža kao izazov digitalnom marketingu

Sažetak

Ogromne količine podataka u obliku poruka na društvenim mrežama izazov su digitalnom marketingu i marketinškoj analizi kako bi se spoznali zahtjevi, potrebe, zadovoljstvo korisnika usluga ili proizvoda. Marketing nastoji biti dio ukupne kulture temeljene na podatcima i definiranju marketinške strategije koje odgovaraju potrošačima, a time i osiguravanju ekonomske koristi samome poduzeću. Stoga se fokus marketinške analize usmjerava i na podatke zabilježene na društvenim mrežama. U radu se prikazuje jedna moguća integracija informacijske tehnologije i *data mining alata* s ciljem vizualiziranja stavova i mišljenja na društvenim mrežama u obliku *word cloud*, a što se potom može uporabiti za oblikovanje marketinške strategije i unaprjeđenja odnosa s kupcima i korisnicima usluga.

Ključne riječi: dubinska analiza teksta, word cloud, marketinška analitika, R jezik, data mining