Analysis of Data from the Social Media

Ladislav Burita, Taha Nejad Falatouri Moghaddam

***Abstract*—**Paper deals with analysis of data from the social media. It starts with explanation of used terms and specification of the social media importance in various areas of business and social life. Authors present two experiments: 1) Sentiment analysis of the Instagram data, 2) Content analysis of the Facebook data. The detail results of the both experiments with comments are included.

***Keywords*—**Sentiment and content analysis, social media, data mining; categorization; statistics

I. Introduction

Predicting of success, the future activities has been always a debating for all businesses [1]. Getting access to first hand data for these prognostications is not an easy process where most of the customers are not available after purchase.

While the social media (SM) could create a new area of investigation on this issue [2]. These days' the SM is an inseparable part of human life's this is more radical for the youngster who already lives on SM [3].

Facebook by having more than 1.79 billion users is the most popular social network, following by Instagram by having more than 500 million daily active users where more than 95 million of photos and videos share daily [4].

The amount of data daily published on social media made a potential opportunity for most business to attend. Where most of the Fortune 500 member have been established their own SM analytics system. It is estimated that the total market of SM from 1.6 billion dollars in 2015 will get 5.4 billion dollars in 2020 [5].

TABLE I
SOURCES FOR THE SOCIAL MEDIA ANALYSIS

Source for SMA	%
Microblogging	46
Review	17
Blogpost	10
Internet forum	9
Review	7
Q&A	6
Tag	5

The use of social media analysis (SMA) is versatile; it has been used in banking, education, child welfare, tourism, marketing, entertainment, government, food industry, clothing, etc. [6].

SM is any web based service with these aspects: First all the members have to sign up as a profile; Second they can make links with each other, and finally the user can share original content, or re-share second hand content there [7]. In the same paper is mentioned the percentage of using sources for the SMA, see Tab. I.

Dell Company proposes the most comprehensive definition for the SMA: "An evolving business discipline that aggregate and analyzes online conversation (industry, competitive,

L. Burita, University of Defence, Brno Czechia (e-mail: ladislav.burita@unob.cz).
T. Nejad Falatouri Moghaddam, Tomas Bata University in Zlin, Zlin Czechia (e-mail: falatouri_moghaddam@utb.cz).

prospect, and customer) and social activity generated by brands across social channels. SMA enable organizations to act on the derived intelligence for business results, improving brand and reputation marketing and sales effectiveness and customer satisfaction" [8].

Authors present two experiments in the paper:

1. SMA of Instagram with the goal to discover research results of the sentiment analysis of the supermarket chain customers.
2. SMA of the Facebook with the goal to discover research results of the content analysis with the key word "military".

II. Sentiment Analysis Experiment

A. Literature Review

By Advent of Web 2.0 technologies, the content providing in the web has been changed from publisher oriented to the user oriented, where programing abilities is not needed to broadcast a content. The personal background develops data in the SM and daily activates [8]. The user generated content is one of the important resource of SM and the valuable content on it [9].

Sentiment analyzes is a called opinion mining and it is based on the feeling of the customers. SMA is widely used in finding out the link and connections. Text mining is a method for extraction information from unstructured data. In the paper [10] is proposed clustering that combines sentiment tone, relevance, keyword analysis, intensity and alert analysis.

Personal comments show the influence of cooperating attendance in the SM. Some research results have been done on the correlation of comments and online shopping, while for the offline shops this correlation has not been investigated by the researchers [11]. Especially relating to lack of access to consumer after purchase experience directly. To overcome this shortage, the researchers tried to utilize SM comments as the useful source of consumer experience [12]. Attending in the SM where the buyers could contact the companies with no mediator, aware the companies of consumer sentiment, and improve brand reputation by gaining follower [13].

B. Research Questions

As was mentioned in the literature review, companies use the SM posts to attract customers. In this study, we are implementing data mining (DM) methods to find out how to improve influence of Instagram reputation of Ofogh-Koroush (OK). To obtain this, we analyze the last 100 post of OK Instagram page to find out a reliable rule, see Fig. 1.

Fig. 1 Ofogh-koorosh Instagram page

Investigating Instagram posts of OK Chain stores, we came across with six types of post sharing (see Fig. 2):

1) **Events:** These posts are related to some special days such as Mother's day. Sometimes the page owner shares some related content for celebrating the event or gathering attention.

These posts are mainly finished by related Hashtags.

2) **Voting:** In these posts, a question is asked about the performance and satisfaction level of the customers about a specific situation.
3) **Competition Result:** The page owner uses this media to announce the winners of any competition and campaigns in the physical stores.
4) **Self-Advertisement:** These are advertisement of OK's services and staffs' performance.
5) **Sale:** Daily sales and discounts are announced via these posts.
6) **Product Advertisement:** OK's third party cooperation's products are advertised in these posts, too.

Fig. 2 Sample of the six types of OK's Instagram posts page

To find out which type of posts are more influential for increasing OK chain stores reputation, we have to answer three main questions.

Q1: Which type of contents could bring more comments? The company has allocate some money for content providing by recognizing the most attractive type of content the money flow could invest in efficient way.

Q2: What is the best time to share a post on Instagram? It is mentioned in the literature that the number of daily posts in Instagram reach 100 million daily while most of the user follow different pages from different countries by choosing the right timing of content sharing the users could not lost your posts.

Q3: Which type of posts bring sense that is more positive? Positive comments could attract more users to follow and improve reputation of the brands. In this case, we need to understand the sentiment of the company user for each category, to emphasis on it.

C. Methodology and Analysis

Majority of comments in this page were in Persian language, therefore we faced several difficulties and challenges and had to set few changes to have a clean usable data set. The Iranian language has official speaking, which is used in the News, is a spoken language. More than 80% of the content was written in spoken language, which may be processed wrongly by the data mining software. In order to increase the accuracy, we set a dictionary to change the spoken verbs to official verbs.

The Uses of emoji in many comments make some barriers for the application to analyze the sentiment. To overcome this problem, we used some queries to change emoji to a word or

sentence based on Iranian culture; for instance, 🙌 to thank you or 😍 to like.

Persian language comment with English characters are rare and hard to be understand by programming but we used Google application (https://www.google.com/intl/fa/inputtools/try/) to overcome it. In some cases, and events such as football match prediction competition the user have to comment their guess of result, which is considered unrecognizable for most of the algorithms. We decided to change the real answer to positive answer and not related comment as a negative answer.

Near 15% of the comments were related to the admin of Instagram page and we eliminate them. In some events, the owner set competitions for mentioning other people. These posts could not be a part of our investigation relatively near to 300 comments was related to mentioning people. It was not recognizable for us to find out if it is a positive reaction or negative. By removing the duplicated comments, the final data set include 3263 comment of latest 100 post of OK Instagram. Example of the data set is at the Fig. 3.

ExportComments.com			
Source URL	https://www.instagram.com/p/BvHBptwAjos/		
Name (click to view profile)	Date	Likes	Comment
mr.ballon24	17/03/19 14:32:32	0	دوستان از پیج من هم دیدن و در صو
bijan_h.ma	17/03/19 14:44:35	1	تخفیف ها رو ول کنید به مشتریان اون
llvllr_r	17/03/19 15:05:04	0	@bijan_h.ma عید شد خبری نشد
hashem.shirazi	17/03/19 15:36:59	0	مفتشم گرونه
damnoshsaraneshaa	17/03/19 15:58:21	0	سلام ی مرغ و گوشت وبرنج وحبوبات
beti6672	17/03/19 17:04:18	0	👏👏👏
3427_maryam	17/03/19 18:38:26	0	بله خریدم از عرقیات عالی بود
hello_hazarat	17/03/19 18:52:29	0	ببخشید این امتیازات توی نرم افزار باش
nova_concept1	17/03/19 22:52:21	0	عرقیجات
rozhman_esmaili	18/03/19 16:02:31	0	یک مینو سفارش میدم
nahal.shafaatian	18/03/19 21:58:01	0	کالاهای اساسی تخفیف بزنید سرکه و
zima_gate	20/03/19 05:04:34	0	خوشم اومد😊👍

Fig. 3 Sample of downloaded comments using exportcomments.com

D. Answering Question

Answering the first question, we count the number of comments of each type of posts; the result is in the Tab. II. It is worth to mention that the company has to allocate some amount of money on content providing. According to the result, it is revealed that the Event's posts are the best for comment gathering, it could mainly be related to the trends and hashtags for that special day.

TABLE II
AVERAGE NUMBER OF COMMENTS FOR EACH POST TYPE

Post Type	Average number of Comments
Event	82.40
Voting	68.60
Competition Result	58.11
Self-Advertisement	51.57
Sale	36.29
Product Advertisement	24.50

The next we investigated how the time of posting comments to find out; what is the best time for publishing a post. The expected sense in this affair was that sharing at night while people are at home is the best choice, though as it is shown in Fig. 4, we find out that most comments have been sent between 2 pm to 7 pm while people are free of work to do shopping. It means

that best time to share a post is by the noon and two or three hours sooner than 2 pm.

The last and most important part of our research is the sentiment analysis of the comments in the six mentioned categories. For this, we used Rapid Miner studio and Rosette text Analytics extension (https://www.rosette.com/rapidminer/) that can support Persian language.

The process of sentiment analysis is shown at the Fig. 5. The results were eliminated to only Positive, Negative and Neutral comments, see the Fig. 6; the most positive comments are received on the events posts.

It shows that investment in these rather activates could worth and the problem is in sale category where most of the customers send their complaint. The least important category is the product advertisement, although the company could benefit from product advertisement by asking for the advertisement fee it is not affect its reputation online.

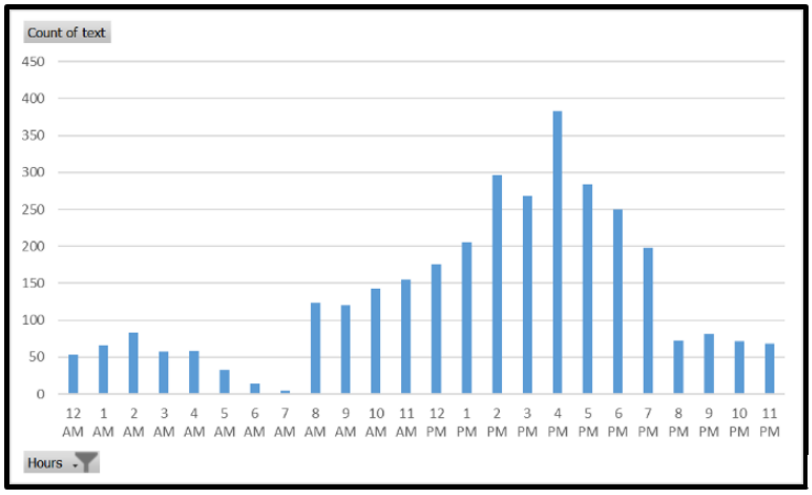

Fig. 4 Number of comments on daily time

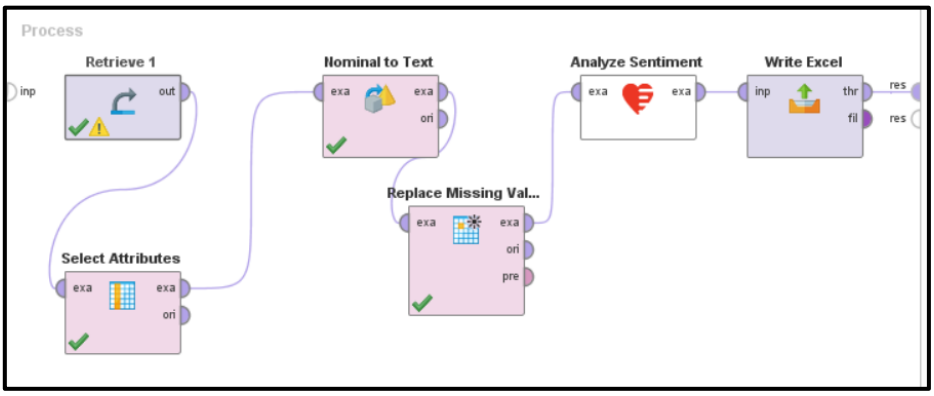

Fig. 5 Process of the sentiment analysis

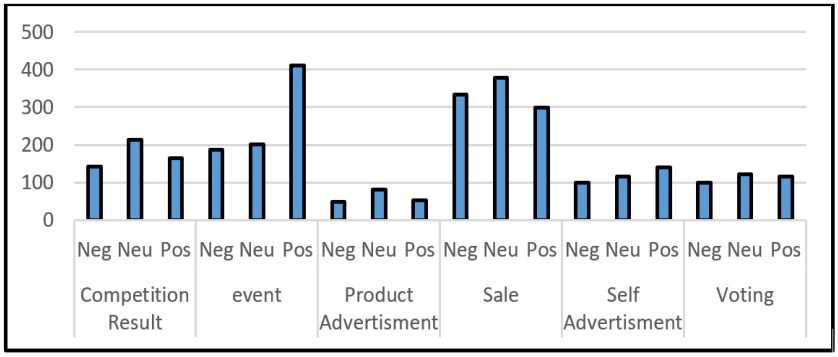

Fig. 6 Process of the sentiment analysis

III. Content Analysis Experiment

A. The Literature Review, Research Question and Hypothesis Development

The literature review was oriented to find the effective tools for data extraction from the Facebook. From the set of tools in [15] was selected only those that can be used without programming. Some tools are using conditional participation in the development team, which is also not appropriate for the authors.

The research experiment was prepared with data, obtained from Facebook, using tool Netvizz [16]. The Netvizz application provides "raw" data for both personal networks and pages, but provides data perspectives not available in other tools, e.g. comment text extraction; it also provides data for groups, a third functional space on Facebook. Running as a Web application, Netvizz does not require the use of Microsoft Excel on Windows like NodeXL and thereby further lowers the threshold to engagement with Facebook's rich data pools [17]

The aim of research [18].was to analyze the content posted by Municipal and State Tourism Organizations (DMO) of the twelve headquarters cities and States of the FIFA 2014 World Cup in their fan pages on Facebook. In the first stage, the official Facebook fan pages were identified, then posts published between June 1st and July 31st of 2013, period from pre to post-event FIFA Confederations Cup Brazil 2013 were collected. The data analysis method employed was content analysis from the perspective of Bardin (2011), which is divided into: i) pre-analysis using dedicated SW, phase ii) material exploration and iii) treatment of results, inference and interpretation. It was observed that the DMOs analyzed publish diversified information to users, including actions addressed to the abovementioned event.

An academic group and discussion forum were established on Facebook for a cohort of postgraduate students studying the concepts and principles of eLearning. The Forum had a constructivist, student-centric ethos, in which students initiated topics for discussion, while the course leader and administrator facilitated. Previous research has been conducted, involving content analysis of the topics and academic discourse, but the present study focuses on social aspects, investigating social-and study-related pursuits and determining whether synergy can exist between them. A literature review shows how social networking by students, initially social, began to overlap with academia, leading to the use of groups for academic purposes and forums for subject-related discussions. In the present study, data was triangulated and two methods of data analysis were used [19].

The research question: Find objects and subjects, services or activities, connected with the key word "military".

The working hypothesis:

H1: Data does not obtain any specific military activity, connected with warfare. Subjects and objects are not in detail described from the military organization point of view.

H2: The most records offers any services for military support or offer sale of any products from the military environment.

B. Data Acquisition and Research Results

In the basic form, 100 records can be obtained in tabular form, the structure of which consists of fields: identifier, name, check-ins, description, cover-picture, link to Facebook, and link to website. After the initial examination were removed duplicated records (4) and records that do not match the keyword query "military" (8). The remaining 88 records were subject to farther analysis. The filled values in records were incomplete, for example description (28%), cover-picture (45%), and link to website (59%). The detailed statistical analysis includes:

- Records by country of origin (Fig. 7).

- Analysis of records by category (Tab III).
- Arrangement by contributor's area of interest (Tab IV).

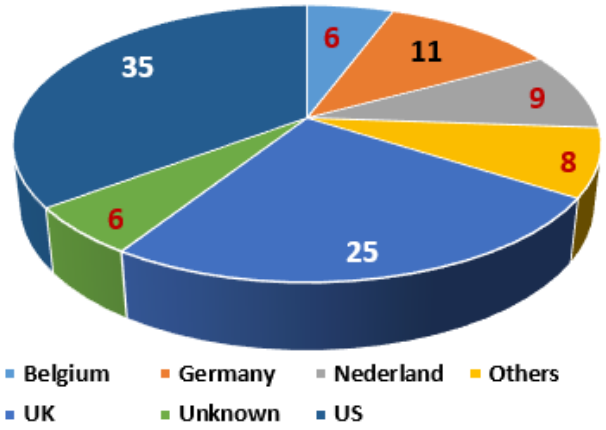

Fig. 7. Records by country of origin (%)

TABLE III
RECORDS BY CATEGORY

Category	No	%
Military shop	26	30
Military service	12	14
Military Academy or College	9	10
Military base or camp	8	9
Military Recruiting Office	6	7
Military cemetry	5	6
Military museum	5	6
Military restaurant	4	5
Military community	3	3
Military charity	2	2
Military press	2	2
Military road	2	2
Military Intelligence	1	1
Military Order	1	1
Military Police	1	1
Military racing	1	1

TABLE IV
RECORDS BY AREA OF INTEREST

Area of interest	No	%
sale	21	24
advertisement	11	13
education	10	11
history	10	11
recruitment	8	9
social	8	9
military installation	5	6
sport	5	6
housing	2	2
topography	2	2
association	1	1
honour	1	1
journal	1	1
military tatoo	1	1
psychology	1	1
publisher	1	1

Records by country of origin, see Fig. 7, is the statistics of number records that were included in any country. The most records comes from US (35), UK (25), and Germany (11); the other countries means France (2), Canada (1), Pakistan (1), Philippines (1), Russia (1), and Spain (1).

Result of records analysis by category, see Tab. IV, shows the military topic and goal of the including information.

C. Hypothesis Verification

The source for evaluation hypothesis H1 was detailed inspection of data. There is only one record about military unit ”The 40th Military Police Battalion” and NO records about military activities, connected with warfare. **The hypothesis H1 is true.**

The source for evaluation hypothesis H2 is Tab. III and for its verification is Tab. IV. Military service, support and sale in the Tab. III includes military shop, service, Recruiting Office, community, and charity, together 56%. Military service, support and sale in the Tab. IV

includes sale, advertisement, recruitment, social, housing, association, and psychology, together 59%. The result is more than 50%. **The hypothesis H2 is slightly true.**

IV. Discussion and Conclusion

The study of sentiment analysis identified the most attractive Instagram context of OK page based on text mining; the sentiment of user comments was analyzed in six different post categories (Events, Voting, Competition Result, Self-Advertisement, Sale, and Product Advertisement) to establish a reliable role for the future patch. The most attractive type of post form are Events and Sale where the event's post received more positive comment and sale post received more negative and neutral comments. According to the research on the time of leaving a comment, the best time to share a post would be between 2-4 pm.

The research results of the data content analysis from Facebook are quite different from the sentiment analysis in the first experiment. The goal of the study is searching objects, subjects, and activities, connected with selected key word. In our case, it was "military". This is useful in inspection of SM participants' interests. The surprising finding was relatively large representation of the military history (museum, cemetery, and order) at the Facebook.

The comparison both experiments summarizes following facts: The data acquisition, data content and format is nearly the same (table). The quality of data was similar, about 10% records was excluded (off topic, duplicity). In the first experiment was used for SMA a data-mining tool; in the second experiment categorization and statistics. The results of the first experiment are useful in marketing and customer satisfaction, and of the second experiment for insight into the SM data.

Acknowledgment

Visegrad Fund

This publication is a result of the project implementation: Exhibition and Special Discussion Section on Info and Digital Technologies; reg. no. 21830315. The project is co-financed by the Governments of Czech Republic, Hungary, Poland and Slovakia through Visegrad Grants from International Visegrad Fund. The mission of the fund is to advance ideas for sustainable regional cooperation in Central Europe.

References

[1] G. Chen, et al., "NPP: A neural popularity prediction model for social media content," Neurocomputing, 2019, p. 221-230.

[2] D. Baum, et al., "The impact of social media campaigns on the success of new product introductions" Journal of Retailing and Consumer Services, 2018.

[3] T.P.S. Humaniora, "83 Percent of Teenagers Inseparable from Social Media," 01/06/2016; Available at: http://news.unair.ac.id/en/ 2016/06/01/83-percent-of-teenagers-inseparable-from-social-media/.

[4] M. Ahlgren, "Top 28 Instagram Statistics & Facts For 2019," Available at: https://www.websitehostingrating.com/instagram-statistics/.

[5] I. Lee, "Social media analytics for enterprises: Typology, methods, and processes," Business Horizons, 2018. 61(2): p. 199-210.

[6] N. Misirlis, and I.M. Vlachopoulou, "Social media metrics and analytics in marketing–S3M: A mapping literature review," 38(1), 2018, p. 270-276.

[7] N.A. Ghani, et al., "Social media big data analytics: A survey," 2018.

[8] Association, I.R.M., "Social media and networking: Concepts, methodologies, tools, and applications," 2015: IGI Global.

[9] X. Xu, et al., "Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors," 2017. 37(6): p. 673-683.

[10] I.J.B.H Lee, "Social media analytics for enterprises: Typology, methods, and processes," 2018. 61(2): p. 199-210.

[11] X. Li, C. Wu, and F. Mai, "The effect of online reviews on product sales: A joint sentiment-topic analysis," Information & Management, 2019. 56(2): p. 172-184.

[12] A. Lawani, et al., "Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston," Regional Science and Urban Economics, 2018.

[13] M. Schaarschmidt, and G. Walsh, "Social media-driven antecedents and consequences of employees' awareness of their impact on corporate reputation," Journal of Business Research, 2018.

[14] A. Erz, B. Marder, and E. Osadchaya, "Hashtags: Motivational drivers, their use, and differences between influencers and followers," Computers in Human Behavior, 2018. 89: p. 48-60.

[15] D. Freelon, "Social media data collection tools," Available at http://dfreelon.org | @dfreelon

[16] Netvizz v1.6 – tools to analyze the Facebook platform. Available at https://apps.facebook.com/107036545989762/

[17] B. Rieder, "Studying Facebook via Data Extraction: The Netvizz Application". Available at http://thepoliticsofsystems.net/permafiles/ rieder_websci.pdf

[18] A.A. Biz, C.K. Santos, E.M. Bettoni, et al." Analysis of content conveyed by the tourism departments of cities and states the headquarters of the World Cup 2014 on your Facebook pages," PASOS-REVISTA DE TURISMO Y PATRIMONIO CULTURAL, 14, 2, 2016, pp. 543-559.

[19] R. de Villiers, M.C. Pretorius, "Academic Group and Forum on Facebook: Social, Serious Studies or Synergy?" PROCEEDINGS OF THE 6TH EUROPEAN CONFERENCE ON INFORMATION MANAGEMENT AND EVALUATION, 2012, p. 63-73.