Text analytics on the case of Macedonian companies

ISSN 1857-9973

657.37:81'373.46]:661.12(497.7)“2016/2018“
657.37:81'373.46]:654(497.7)“2016/2018“
657.37:81'373.46]:336.713(497.7)“2016/2018“

Dushko Todevski*[1]*,

[1]Faculty of Economics, Goce Delcev University- Stip, Republic of Macedonia

duskotodevski@gmail.com

Abstract

The subject of this paper is to present the content analysis with some of the latest text analytics tools on the case of three Macedonian companies Alkaloid, NLB Banka and Makedonski Telekom. The subject of text analytics is annual addressing of the CEO’s (Chief Execute Officer) integral parts of the annual reports for 2016, 2017 and 2018. Annual reports are used as a text input for the text analytics case. The text analytics case is presented through bag of words and word cloud for determination of key terms frequency. The frequency of the key word (tokens) selected in each category indicates the focus of the company for the period addressed. At the same time this frequency status suggests the focus of the company in the future involvements.

Sentiment analysis included in this text analytics case aims to determine the opinion expressed by the CEO’s related to the reported and following business period. At the same time feature of topic modeling is extracting topics from the CEO’s addressing for each company for the observed period.

The goal of this paper is to present the analytical framework for business content analysis and to test it on the text input in the example of the annual reports. In our case results and findings from the analysis suggest that three companies through this analytical framework present different management strategies with different focus on operational and market segment. Alkaloid according to the result has orientation on maximization of management performance and decision making. The focus on NLB bank is on the continuous improvement on its competitive advantage against competition. Although, the case of Makedonski Telekom with the presented results suggested that the company focuses on the continuous improvement of services, products and the network infrastructure supporting them.

Key words: text analytics, companies annual reports, sentiment analysis, bag of words, topic modelling

1. Introduction

Content analysis generally suggest the qualitative analysis of the content of the observed text [1]. In this paper we suggest another type of qualitative analysis to be use for determining the messages, content, narrative, opinion and logic declared in the annual reports. Framework used in this paper will provide other perspective for analysis than classical determining the meaning, narrative, or the messages from the text published in annual reports. The suggested analytical approach should offer more objective analytical insight on subjective statements from content. This text analytics approach applied in different business segments and cases can help with creation of other variables for further quantitative analysis. The use of this framework is recommended for use alongside with classical contents analysis, validating the meaning, messages or narrative from the text input [2].

Text analytics framework used in this paper uses machine learning algorithms applied in techniques such bag of words, word cloud, sentiment analysis and topic modeling [3]. All this algorithm use python code in background for calculations. In this case we use text input of several text articles, although these algorithms make best performance on big text data [4]. All calculations in this paper are made with Orange3 – graphical user interface of the popular orange library written in python code [5].

Presented recommendation from this use case also can be used simultaneously with classical fundamental analysis of the company or can be used for creation of proxy variables for more complex machine learning cases.

2. Data and Methodology

The data for this text analytics case was collected from the annual reports of the Alkaloid, NLB Banka and Makedonski Telekom [6],[7],[8],[9], [10],[11],[12],[13]. The primary phases included more observed companies but due to availability of the reports and address by the CEO, the final list included only this three companies. The selected companies are also listed in Macedonian Securities exchange commission and have one of the most liquid stocks in the Macedonian market. Alkaloid is a pharmaceutical company operating on global market, with the focus on European market. NLB Banka is one of the largest banks on Macedonian market by assets and market share, while Makedonski Telekom is one of the two telecommunication operators on Macedonian market. Data – related text was collected from the annual reports downloaded from the corporate web sites of the Alkaloid, NLB Banka and Makedonski Telekom.

The annual reports as a data sample included address by the CEO's from Alkaloid, NLB Banka and Makedonski Telekom for the period of the year 2016-2018. The CEO for Alkaloid and NLB Banka were Zhivko Mukaetov and Antonio Argir, while CEO for 2016 for Makedonski Telekom was Andreas Maierhofer, for 2017 was Andreas Elsner, and for the 2018 was Nikola Ljushev.

Table 1 CEO Address by company and year

Company	Year	Total words count
Alkaloid	2016	1075
Alkaloid	2017	1035
Alkaloid	2018	1203
Makedonski Telekom	2016	567
Makedonski Telekom	2017	756
Makedonski Telekom	2018	723
NLB Banka	2016	1208
NLB Banka	2017	2190
NLB Banka	2018	1627

Source: Authors calculations

The main assumption of this analysis is that the statements in the CEO addressing reflects the real facts from the referent year. Another important assumption of this analysis is that the strategies, goals, and plans presented in the CEO address for the following period are corresponding with the company's real business potential.

Primary hypothesis of this paper is the difference in the approach of the several categories as **customers- clients, products or services, shareholders, operations and market**. Below

in the part of Bag of word and word cloud will be presented ranking of the words represented in each category and in each company. This difference on the ranking should be evaluated with the help of the bag of words and word cloud.
The analysis part of this paper will also include sentiment analysis and topic modeling as the text analytics techniques for this qualitative analysis.

3. Bag of words and word cloud

Bag of words as a text analytics tool in this analysis is used for terms frequency count, while the word cloud is used for visual interpretation of the bag of words results [14]. Basically, with the bag of word we will have all frequency count of all words used in the CEO addressing. The word cloud visualization uses the frequency and presents all words connecting the frequency with the front size and color of the words in the cloud form presentation [15].

Below are the visualizations of the bag of words with the word cloud for the three companies:

Figure 1 Word cloud Alkaloid

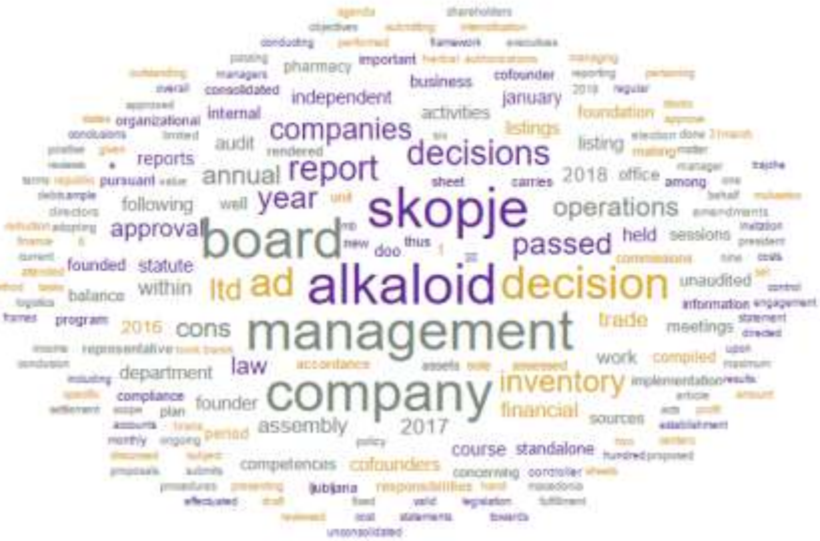


Figure 2 Word cloud Makedonski Telekom

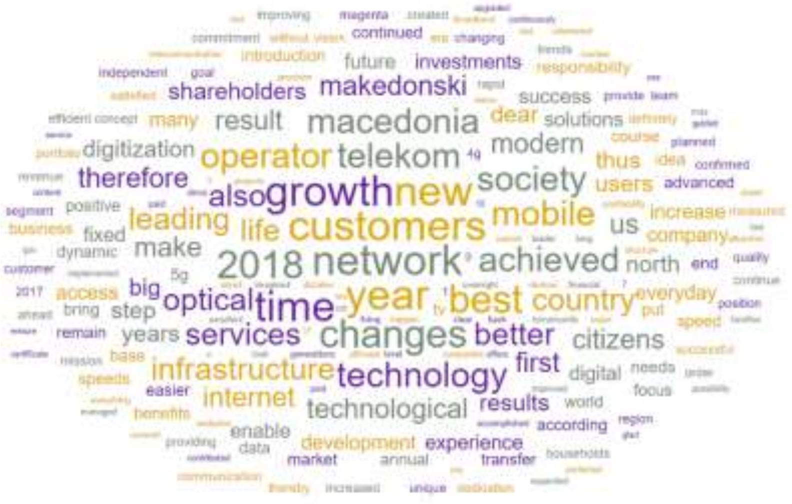
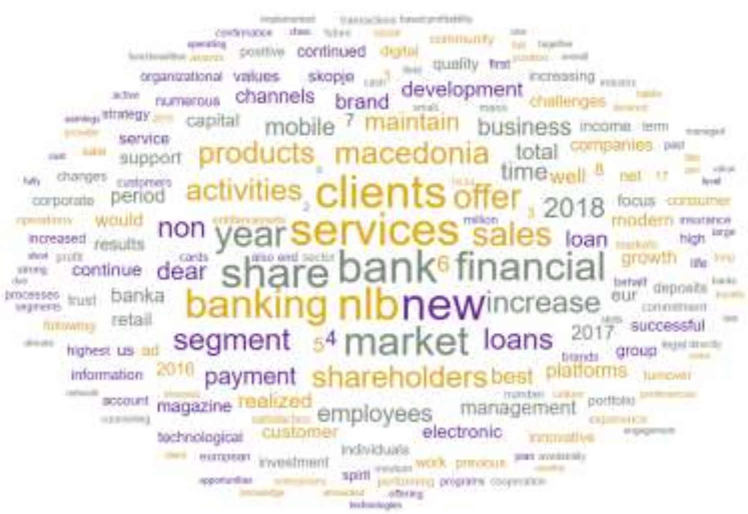


Figure 3 Word cloud NLB Banka



In the table 2 are presented results from the top 20 most frequented terms used in the annual address included in the company's reports. Some of the words are marked with asterix, which represents the affiliation to the group represented with the asterix. The idea here is to determine the terms by categories which are presented in the text. The frequency of the terms and their categorization should reflect the company's focus in the customers- clients, products or services, shareholders or operations and market segment. The categories used in this analysis are not driven with formal criteria and it is more like a rule of thumb criteria for the categorization of the key terms. However, if higher precision is needed i.e. in similar analysis the formal use of criteria for categorization is recommended.

Table 2 Comparation of Bag of words term in three cases for top 20 terms

	Alkaloid		Makedonski Telekom		NLB Banka	
	word	frequency	word	frequency	word	frequency
1	alkaloid	90	network**	19	bank	71
2	management***	67	year	18	services**	41
3	skopje	67	customers*	18	nlb	38
4	board***	61	new***	12	clients*	38
5	company***	57	telekom	10	new**	32
6	decision***	45	macedonia	10	share***	30
7	ad	38	services**	9	market***	28
8	decisions***	31	growth***	9	banking	28
9	report	31	best***	9	financial	28
10	inventory***	27	changes***	8	year	25
11	ltd	27	makedonski	7	sales***	23
12	passed	27	result***	7	offer**	23
13	year	27	us	7	macedonia	21
14	companies***	25	digital**	7	shareholders****	20
15	operations***	22	company***	7	activities***	20
16	annual	21	also	7	increase***	20
17	cons	21	infrastructure**	7	segment**	20

18	approval***	19	mobile**	7	products**	20
19	law	19	make	7	loans***	19
20	2017	16	society	7	non	19

Source: Authors calculations

Terms including or indicating participation to the following categories :

* **customers- clients,**

** **products or services**

*** **operations and market**

****** shareholders**

According to results top 10 most frequently used words in the case of Alkaloid are: alkaloid management***, skopje, board***, company***, decision***, ad (from Alkaloid AD), decisions***, report and inventory***. Analyzing top 20 most frequent terms by predefined categories the terms we can see that 9/20 terms belong to category operations and market.
In the case of Makedonski Telekom top 10 most frequently used terms are: network****, year customers*, new***, telekom, macedonia, services**, growth***, best***, changes***. By categories labeling of the terms 6/20 word belong to the category operations and market, 5/20 belong to the category of products or services, 1/20 in customers- clients.
Case of NLB Banka in the top 10 most frequently used words has the following list of terms: bank, services**, nlb, clients*, new**, share***, market***, banking, financial, year. Looking at the top 20 most frequent terms by category we have the following result: 6/20 word belong to the category operations and market, 5/20 belong to the category of products or services, 1/20 in customers- clients, and 1/ 20 in the category of shareholders.
The terms contained in the companies' official name or the words suggesting location of the company were not included in any category. Also, all terms because of text preprocessing setup are used with lower cases.
The analysis of the terms and their belongings in the predefined categories showed that there is significant difference in the treatment of the customers- clients, products or services, operations and market and shareholders. Looking from this perspective we can conclude that the primary hypothesis (difference in the approach to different business categories) is confirmed and at the same time presents the source of the differentiation.

4. Sentiment Analysis

Sentiment analysis is an analytical tool usually performed by some python programing language library. The goal is to present objective results from the analysis on the subjective opinion of the author's expressions included in the text [16],[17]. Sentiment analysis in one of its types makes score on words with positive and negative sentiment. In this type of sentiment analysis calculation is made on the summary of the positive and negative scores and the final score is a sentiment score of the text. In our case we will use this type of sentiment analysis with summary score [8]. Another type of sentiment analysis creates separate scores and evaluation of positive, negative and neutral terms used in the input text [17].

Table 3 Sentiment scores on the case of Alkaloid, Mekedonski Telekom and NLB Banka for the period 2016-2018

Report	Sentiment Score
Alkaloid 2016	1.35243
Alkaloid 2017	1.4876

Alkaloid 2018	1.5919
Alkaloid 2016-2018	1.4809
Makedonski Telekom 2016	5.94679
Makedonski Telekom 2017	5.18868
Makedonski Telekom 2018	4.88998
Makedonski Telekom 2016-2018	4.99195
NLB Banka 2016	3.64807
NLB Banka 2017	4.49483
NLB Banka 2018	2.97082
NLB Banka 2016-2018	3.79507

Source: Authors calculations

Sentiment analysis Alkaloid with the scores of 1.35 in 2016,1.49 in 2017 and 1.59 in 2018, represents the author positive outlook for the company in the observed period. The positive uptrend indicates the more optimistic views and expectation of the company for the future period.

Sentiment score in the case of Makedonski Telekom presents the downtrend in the sentiment score during the period 2016-2018. Sentiment score starts with 5.95 in 2016, moving to 5.19 in 2017 and ending at 4.89 in 2018. Highest registered sentiment score in 2016 and significantly lower values in 2017 and 2018 present significant decrease in the expectations of the companies. Despite the high values of the sentiment score and high average score for the 2016-2018 period the downtrend indicates of the lowering the positive expectations of the CEO's related to the observed and future period expressed in their annual addressing.

Sentiment score in the case of NLB Banka presents the values of 3.65 in 2016, 4.49 in 2017 and 2.97 in 2018. The expressed opinion if concluded from the sentiment score indicates that the most optimistic year for the company was 2017. According to the score the most pessimistic year was 2018. Besides the fact that the scores are positive and suggest positive opinion the lowest value of registered sentiment score of 2.97 indicate the lowering the positive perception for the NLB Banka for the year 2018 and for the following period.

Performing comparative analysis of the sentiment score in these three cases where we have the companies from different industry background can be highly biased. However, for some future explorations the comparation of the sentiment scores between the companies from the same industry and on the same geographical market is highly recommended and it should provide valuable and detailed insight about observed business outlook.

5. Topic modeling

Topic modeling as text analytics tool is creating topics, analyzing the structure of the observed and text. With the help of programing language, coding and text analytics libraries, topic modeling as text analytics tool creates abstract topics based on the word's frequency clusters. This foundation of this technique is creating of statistical models from the similarities in the text sample and presented them as separate abstract topics [18]. There are few settings of the topic modelling in Orange3 and in this case we will use Latent Dirichlet Allocation with creation of three topics [19]. Below are the results of each of the three created topics with their key terms for each of the companies.

Topic modeling Alkaloid

1. alkaloid, management, company, decision, board, skopje, decisions, report, passed, ad
2. alkaloid, management, board, company, skopje, decision, ad, ltd, passed, inventory
3. alkaloid, skopje, management, board, company, ad, decisions, decision, report, companies

Key terms present in each of the topic are management, board, decision and company.
According to the three topics Alkaloid and having in mind the most common terms in all three topics, the conclusion is that the company is oriented and believes into the company's generic growth based on good management practice with good decision making.

Topic modeling Makedonski Telekom

1. network, year, customers, macedonia, technology, operator, growth, 2018, life, also
2. network, year, customers, new, growth, changes, time, mobile, best, society
3. network, growth, time, customers, year, best, new, 2018, society, telekom

Key terms present in each of the topic are network, customers and growth.
The results from the topic modeling and the common key terms present in all three topics indicate that Makedonski Telekom believe in generic growth driven by its network and customer growth.

Topic modeling NLB Banka

1: bank, services, clients, nlb, new, share, market, offer, banking, increase
2: bank, nlb, share, market, financial, clients, macedonia, sales, new, services
3: bank, clients, nlb, new, services, banking, financial, year, share, market

Key terms present in each of the topic are services, clients, market and share.
According to the topic modeling results and the common terms in all three topics we find strong believe in competitiveness and complete orientation to the gaining more competitive advantages and stronger market share.

According to the key term in each category in overall, the three companies threat different topics in their CEO address integrated in the annual report. Alkaloid is oriented to management performance, Makedonski Telekom in generic growth driven by the network and customer growth and NLB completely oriented to the market positioning and improving its competitive advantage.

6. Conclusion

The goal of this paper was to provide evidence of the difference between the views presented in the annual reports, specifically in the CEO address as integral part of the annual report. The evidence to support the main hypothesis of different views was offered in all three text analytics techniques. Bag of word and word cloud analysis as well as the key term by category presented the case that there is a significantly different key terms in the presence in the business categories labeled as customers- clients, products or services, shareholders, operations and market.
Sentiment analysis offered results which presented different opinions and views for the period of 2016-2018. Despite the positive scores, this analysis presented decrease of positive opinion in the case of Makedonski Telekom, increased optimism in the case of Alkaloid. The sentiment analysis in the case of NLB Banka presented mixed values of positive opinions and for the last year of 2018 presented significant decrease of positive perception. All of this again support the thesis of significant difference in the content, meaning, opinion and its source in the case of this three companies.
Topic modelling with the three determined topics and the common words detected for each company, one more time presented different business approach for each company. Therefore, we have results indicating that Alkaloid is completely oriented to management performance, while Makedonski Telekom believes in generic growth driven by the network and customer

growth. At the same time NLB is completely devoted to the market positioning and improving its competitive advantage.

References

[1] Mayring P. Qualitative content analysis. A companion to qualitative research. 2004 Apr 21;1:159-76.
[2] Hu X, Liu H. Text analytics in social media. InMining text data 2012 (pp. 385-414). Springer, Boston, MA.
[3] Ittoo A, van den Bosch A. Text analytics in industry: Challenges, desiderata and trends. Computers in Industry. 2016 May 1;78:96-107.
[4] Khan Z, Vorley T. Big data text analytics: an enabler of knowledge management. Journal of Knowledge Management. 2017 Feb 13.
[5] Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Štajdohar M. Orange: data mining toolbox in Python. The Journal of Machine Learning Research. 2013 Jan 1;14(1):2349-53.
[6] http://alkaloid.com.mk/content/Pdf/zainvestitori/2018/Audited%20stand%20alone%20financial%20report% 202018.pdf, access date 10.02.2020
[7] http://alkaloid.com.mk/content/Pdf/za-investitori/2018/Audited%20stand%20alone%20financial%20report%202017.pdf, access date 10.02.2020
[8] http://alkaloid.com.mk/content/Pdf/za-investitori/2017/Audited%20stand%20alone%20financial%20report%202016.pdf, access date 10.02.2020
[9] https://nlb.mk/CMS/Upload/Dokumenti/godishni_izveshtai/NLB_2018_Godishen_Izveshtaj_eng.pdf, access date 12.02.2020
[10] https://nlb.mk/CMS/Upload/Dokumenti/godishni_izveshtai/NLB%20Annual%20Report%202017-Eng.pdf, access date 12.02.2020
https://nlb.mk/CMS/Upload/Dokumenti/godishni_izveshtai/2016_eng.pdf, access date 12.02.2020
[11] https://www.telekom.mk/content/pdf/ANNUAL_REPORT_2018_MAKEDONSKI_TELEKOM .pdf, access date 08.02.2020
[12] https://www.telekom.mk/content/pdf/Annual_Report%202017.pdf, access date 08.02. 2020
[13] https://www.telekom.mk/download/TELEKOM_GI_2016/anual_report_2016_EN.html, access date 08.02.2020
[14] Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics. 2010 Dec 1;1(1-4):43-52.
[15] Heimerl F, Lohmann S, Lange S, Ertl T. Word cloud explorer: Text analytics based on word clouds. In2014 47th Hawaii International Conference on System Sciences 2014 Jan 6 (pp. 1833-1842). IEEE.
[16] Liu B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. 2012 May 22;5(1):1-67.
[17] Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. InEighth international AAAI conference on weblogs and social media 2014 May 16.
[18] Wallach HM. Topic modeling: beyond bag-of-words. InProceedings of the 23rd international conference on Machine learning 2006 Jun 25 (pp. 977-984).
[19] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.