Fenyves, V., Tarnóczi, T., & Orbán, I. (2024). Application of text mining in analysing notes to financial statements: A Hungarian case. *Journal of International Studies, 17*(3), 205-225. doi:10.14254/2071-8330.2024/17-3/11

Journal of International Studies

Centre of Sociological Research

Scientific Papers

Application of text mining in analysing notes to financial statements: A Hungarian case

Veronika Fenyves
Faculty of Economics and Business,
University of Debrecen,
Debrecen, Hungary
fenyves.veronika@econ.unideb.hu
ORCID 0000-0002-8737-0666

Tibor Tarnóczi
Faculty of Economics and Business,
University of Debrecen,
Debrecen, Hungary
tarnoczi.tibor@econ.unideb.hu
ORCID 0000-0002-5655-6871

Ildikó Orbán
Faculty of Economics and Business,
University of Debrecen,
Debrecen, Hungary
orban.ildiko@econ.unideb.hu
ORCID 0000-0001-7783-2201

Abstract. Company stakeholders must have reliable and accurate information about the companies falling into their sphere of interest. In Hungary, one of the key sources of information for company stakeholders is the financial statements and related explanations, which are included in the notes of the financial statements (notes). This study used text mining to analyse the Hungarian annual financial statements notes for 2017, 2019 and 2021. The selection of the notes was based on the proportions of each sector in the national economy. The research analysed 28,700 company notes annually, totalling 86,100 documents for the three years. The text mining and generation of the Term Frequency Matrix have performed 'quanteda' packages of the R statistical system, which incorporate the results of artificial intelligence research to enhance the efficiency of text mining. Based on the results, the contents of the notes to the financial statements appear to be a rather mixed picture in Hungary. Analysing the term frequency matrix for the 67 most common terms has revealed no significant difference between the years. However, considerable differences have been caused by size categories and

Received:
November, 2023
1st Revision:
August, 2024
Accepted:
September, 2024

DOI:
10.14254/2071-8330.2024/17-3/11

sectors. The notes are statistically significant using Jaccard similarity analysis, considering the year, corporate size, and sector.

Keywords: notes to financial statement, text mining, text analysis, Hungary

JEL Classification: M41, M42, C12, C63

1. INTRODUCTION

Business organisations make their decisions in an uncertain environment. Management of the effects of an uncertain environment requires an adequate risk assessment and well-informed decision-making provided appropriate and essential information at the designated place, level, and time. Accounting information is one of the decisive elements for financial decision-making. Accounting information comprises data observed and processed previously within a regulated system, providing information on the states and processes of the company's activities in numerical and textual forms. Adequate accounting information allows us to examine trends and determine reasons for time-related differences. This information enables stakeholders to decide based on expected future values linked to past knowledge.

Accounting information has a dual role. First, it must support managerial decision-making and show the consequences of decisions and actions. Second, the published financial statements present companies to their social and economic environment. Accurately interpreting the content of the two main parts of the financial statements (balance sheet and income statement) requires explanatory texts (notes) for their essential components. According to the Hungarian Accounting Act, notes are a textual document attached to the annual financial statement of business companies. The notes must contain all the information, analyses, and explanations necessary for understanding and usability of the data in the financial statements.

The authors started analysing the notes to financial statements because very few studies examined them in Hungary. Several authors investigated the necessary content and importance of the notes (Tóthné Szabó, 2010; Filyó, 2014; Kántor, 2016; Kerezsi, 2017), and analyses were also prepared using a smaller number of companies (Kerezsi, 2021; Kerezsi et al., 2019). However, the number of studies in which text mining would have been used for the investigations and examined a larger sample of companies was only with one author (Kerezsi, 2020). At the same time, many articles related to the analysis of notes to financial statements can be found in the international literature presented in the literature review.

The main aim of the research was to analyse the notes to financial statements in the sample to what extent they provide information to their users. Text mining was chosen as the method of analysis to generate the term frequency matrix. The term frequency matrix was used to answer the questions related to the main aim. The research questions formulated based on the main objective were:

- What are the most frequently occurring terms in the term frequency matrix?
- Are there statistically significant differences between the three years for the most frequently used terms?
- Are there statistically significant differences in the frequency of the most frequent terms across sectors and company size?
- How does the similarity index of the notes change when pair-wise comparing them based on size and sector?

2. LITERATURE REVIEW

2.1. Financial statements as the source of information

As the economy constantly develops and transforms, understanding the consequences of managerial decisions is becoming increasingly essential. The results of these decisions could be seen in the market as the company changes, grows, or declines, but these effects are somewhat subjective, often less visible and comparable. In current conditions, financial statements are the most comprehensive, objective, and reliable information bases (Thalassinos & Liapis, 2014). The main objective of financial statements is to deliver information about the financial position, performance, and changes in the financial position of a business entity for market participants. The financial statements usually contain a balance sheet, an income statement, a cash flow statement, and notes explaining the figures of the previous documents (Lepadatu & Pirnau, 2009; Bohusova et l., 2022; Shakatreh et al., 2023).

Transparent and valuable information about market participants and their transactions is crucial for the market to work efficiently, which is a primary requirement of market discipline. In any well-performing economy, regulated disclosure requirements impact the quality and quantity of the provided information and ensure the market's stability (Lepadatu & Pirnau, 2009). According to Lepadatu and Pirnau (2009), the main attributes of financial statements are transparency, accountability, relevance, reliability, comparability, understandability, and timeliness. The authors highlight that the first two attributes of market stability are paramount. Transparency means that significant information on the firm's condition and operations is accessible to company stakeholders. Moreover, accountability refers to companies needing to account for their actions and take responsibility for their decisions and consequences.

Computerised enterprise management systems have radically increased the information available for analysts and managers, which poses an increasingly severe challenge to decision-makers. The information processing time is rising due to significantly increased information volume, which may limit the market participants from fully processing all the information required for their decision-making (Bai et al., 2019). Osadchy et al. (2018) list the stakeholders of financial statements. They classified the shareholders and lenders of the firm, governmental bodies, customers, suppliers, associations, trade unions, and even the press and information agencies in the group of external parties concerned. They featured the company management and employees as internal stakeholders. As internal information users, management can access more data beyond published information. However, external users can use only the published annual reports and financial statements to support their decision-making (Pakšiová & Oriskóová, 2020). Besides being the basis for making managerial decisions, financial statements can be used to analyse the company's activities and detect the causes of deviations from the standard values. Furthermore, statistical organisations can use financial statements to analyse and forecast the direction and level of the economy's development (Osadchy et al., 2018, Földvári and Erdey, 2009, Bíró et al. 2019).

Another extensively researched field for financial statement analysis is accounting manipulation, which is conscious misinformation by the company's financial statements disclosed publicly. The intention is to mislead stakeholders regarding the firm's financial position by overstating its expectations of assets or understating exposure to liabilities (Jofre & Gerlach, 2018). The accounting profession, the business communities, and regulators are responsible for preventing the manipulation of financial statements. Traditional auditing procedures are not always capable of preventing manipulation, but analysing financial statements using advanced techniques can be an effective tool. Data and text mining can be such an effective tool for uncovering manipulations. Analysts can use former manipulations to build models to help reveal manipulated financial reporting (Gupta & Gill, 2012).

The study conducted by Aymen et al. (2018) analysed how financial information's readability impacts financial analysts' behaviour in France. The researchers examined 88 companies listed on the French Stock

Exchange from 2009 to 2014. They found that improved readability reduces agency costs and information asymmetry, making it more appealing to financial analysts and reinforcing the assumption of adverse selection. In a study conducted by Abernathy et al. (2018), the researchers also analysed how the readability of firms' annual financial statements affects their borrowing costs. The study revealed that the shareholders of companies with less readable financial statements have more difficulty accessing transparent information, leading to higher external financing costs.

2.2. Notes to financial statements

A company's public disclosures contain an essential source of textual information, such as notes to financial statements. They are crucial information originating from inside the organisation. The analysis of this qualitative information supports understanding the meaning behind quantitative details of the financial statements to obtain helpful information about a company's future performance (Kearney & Liu, 2014).

The critical step was the readability research, which resulted in the development of formulas by Flesch in 1940. The Flesch formula enabled the estimation of the intelligibility of written notes without needing to read them. Barnett and Leoffler (1979) used the Flesch Reading Ease Formula to examine the readability of selected accounting and auditing notes in annual statements. Their measurements concluded that notes related to financial statements and independent auditors' reports are at unsatisfactory difficulty levels. Additionally, their analysis indicated that the readability level of current notes to financial statements is significantly lower than the statements in 1969.

2.3. The application of text mining in accounting

The quantitative data obtained from financial statements are usually analysed using traditional statistical methods or data mining techniques. However, textual data, such as notes to financial statements, require advanced analysis solutions for extracting and organising relevant knowledge.

The need to analyse text files arose more than 50 years ago in economics. Stone et al. (1966) wrote about the importance of analysing linguistic texts in economics. The authors defined textual content analysis as a method that allows conclusions by identifying features found in textual material. However, the technical conditions that became available with the development of computer technology were still lacking in performing this type of analysis.

Amani et al. (2017) explored the applications of data mining techniques in accounting. They created a framework to organise the literature on data mining applications in accounting. The framework encapsulates a taxonomy of data mining applications in accounting. It presents a holistic view of the literature and systematically organises it structurally, logically, and thematically coherently.

According to Chan and Franklin (2011), textual data contains more information than numeric data because the former allows us to predict financial trends and justify the predictions. As examples, the authors mention words or phrases like "shortfall" or "resignation", which could appear in the company's statements and can be assumed to expect the company's decline.

Qualitative, textual information comes mainly from corporate disclosures, media articles, and Internet messages. A company's disclosures are the primary source of textual information. The analysis of this qualitative information aims to understand the meaning behind the text style, tone, and frequency of words and phrases, thus obtaining helpful information about the firm's performance (Kearney & Liu, 2014).

Currently, textual analysis, a multidisciplinary field of study, is relatively rare in finance and accounting, but it has great potential due to the volume of documents (audit reports, corporate social reports, management reports). The growth in the use of digital tools and social media increased the volume of non-structured documents available on the internet. Textual analysis is closely connected to computational

linguistics, natural language processing, and content analysis, and it refers to the research activity to extract information from textual sources (Gandía & Huguet, 2021; Yadav and Sora (2021)).

Textual analysis studies can be divided into "studies that examine the readability of text and studies that deal with the extraction of information from the text, in the form of either keywords or targeted terms (Loughran & McDonald, 2016)".

Fenyves et al. (2019) examined how the Hungarian information services sector companies fulfil their obligation to provide needed information prescribed in the Accounting Act in their notes to the financial statements. They used notes to the financial statements of 8,226 companies regulated by the Hungarian Accounting Act for the analyses, which have information technology services as their primary business activity. The research was performed using the text mining method, utilising every available note to the financial statement of the sector. The study's findings reveal that the amount of published information shows greater and lesser differences. In many cases, the quantity of published data does not fulfil even the minimal obligations stipulated legally.

Senave et al. (2023) noted that text-mining techniques have become popular across various industries for extracting knowledge and qualitative measures from textual data. In accounting, text-mining outputs can complement and validate quantitative data. Their study presents an overview of the applications of text mining in accounting practice, including a critical assessment of the typical text mining process, contemporary text mining techniques, and the information obtained through these techniques.

3. RESEARCH DATABASE AND METHODOLOGIES

3.1. Database for research

The research used notes of the financial statements (hereafter notes) from 2017, 2019, and 2021. These notes were purchased separately each year from Opten Informatikai Kft. 10% of the operating companies reported by the Hungarian Central Statistics Office (KSH) were chosen to determine the number of companies included in the analysis. The KSH reported 351,000 business companies in Hungary in 2017. Since financial statements are only prepared by business companies, the researchers selected 10% (~35,000) of operating companies as the analysis database. Opten Informatikai Kft was responsible for choosing the sample database based on the researchers' selection conditions given.

The notes were selected using the following conditions researchers gave. Firstly, companies with a revenue of at least HUF 10 million (equivalent to EUR 32,200 as per the exchange rate of 29 December, 2017) were chosen. Secondly, the total sample was divided among 15 sectors of the national economy (refer to the notes of Figure 1), with their proportion in the sample corresponding to the actual proportions of the national economy sectors. Thirdly, companies within the sectors were categorised based on employee number, as per Table 1. Fourthly, companies within each group were arranged in descending order of sales revenue and divided into quartiles. Finally, a predetermined number of companies was selected from each quartile.

Opten Kft gave notes of 36,841 companies for 2017. Notes for 2019 and 2021 were purchased considering the same companies as for 2017 in the following year. Only those documents were included in the final analysis database, which existed and was processable in both years. Thus, 28,592 documents were analysed in both years.

The study only analysed the notes to financial statements every two years (in 2017, 2019, 2021). After analysing 2017, it was decided that notes would only be purchased every two years, for two reasons. Firstly, it was thought that the changes would be better presented if the following year were not examined. Secondly, purchasing the notes for each year would have required a significant amount, and the financial resources available for research at the institute are limited.

The year 2023 has not been included in the study because companies had to submit their financial reports by 31 May 2024, and a relatively complete database will only be available at the end of this year or the beginning of 2025. The Opten Informatikai Ltd. purchases the database from the Ministry of Justice. Many Hungarian companies upload their financial statements to the e-beszámoló platform late. We cannot purchase the database from the Ministry of Justice because they do not filter the data, which is necessary to ensure the same companies are available yearly. Without this filtering, making comparisons could be problematic.

Table 1 shows that almost half of the companies belong to the second employee category (46.57%). The sum of categories 1, 3, and 4 (50.01%) barely exceeds the number of companies belonging to category 2, and there are no significant differences in the number of companies in these categories. The fewest companies belong to category 5, meaning only a few large companies are in the sample set.

Table 1

Distribution of companies analysed according to employee number categories

Categories	Lower limit of employee number	Upper limit of employee number	Number of companies in the category	Distribution of companies by category
1	0 or not specified	-	4,486	15.69%
2	1	4	13,316	46.57%
3	5	9	5,042	17.63%
4	10	49	4,770	16.68%
5	50		978	3.42%
Total			**28,592**	100.00%

Note: The first category includes enterprises with no contribution-paying employees or that have not reported the number of employees in their financial statements uploaded to e-beszamolo.
Source: own editing

The companies are divided into 15 sectors (A, C, F-N, P-S) (Table 2), which is the same as the European NACE (Statistical Classification of Economic Activities in the European Community, 2008) Rev.2. The NACE Rev. 2. classifies business companies into 19 sectors (A - S). However, four sectors (B - Mining and quarrying; D - Electricity, gas, steam, and air conditioning supply; E - Water supply; sewerage, waste management, and remediation activities; O - Public administration and defence; compulsory social security) were excluded from the analysis because of the small number of companies.

The analysis (Table 2) shows that the four biggest sectors (Wholesale and retail trade, repair of motor vehicles and motorcycles; Professional, scientific, and technical activities; Manufacturing; Construction) represent nearly 62% of the examined companies. The other 11 sectors make up the remaining 38%. The Wholesale and retail trade, repair of motor vehicles and motorcycles sector has the highest number of companies, accounting for more than a quarter (27.83%) of the total.

Table 2

Distribution of companies among sectors

National economic sector codes and names		Number of companies	Distribution of companies among sectors
A	Agriculture, forestry, and fishing	989	3.46%
C	Manufacturing	3,207	11.22%
F	Construction	3,131	10.95%
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	7,957	27.83%
H	Transportation and storage	1,271	4.45%
I	Accommodation and food service activities	1,334	4.67%
J	Information and communication	1,381	4.83%
K	Financial and insurance activities	407	1.42%
L	Real estate activities	1,931	6.75%
M	Professional, scientific, and technical activities	3,401	11.89%
N	Administrative and support service activities	1,279	4.47%
P	Education	256	0.90%
Q	Human health and social work activities	1,378	4.82%
R	Arts, entertainment, and recreation	366	1.28%
S	Other service activities	304	1.06%
Total number of companies		28,592	100.00%

Source: own editing

3.1. Text mining

The quantitative database was created using different functions of two text mining packages of the R statistical system, namely 'tm' and 'quanteda' (Feinerer et al., 2008). The notes were available in PDF format and grouped by sector and size. The 'quanteda' package allows the simultaneous reading and processing of several text files, making it useful for natural language processing (NLP) and text analysis. The AI-powered capabilities of this package make it stand out from other applications, significantly reducing processing time by filtering out filler words and characters and searching for specified terms. Though using 'quanteda' requires knowledge of R programming, but it enables robust and efficient analysis. The 'quanteda' package includes new tokenise rules, making it more efficient and consistent with more languages. The 'quanteda' R package and its linked packages' quanteda.textstats' and 'quanteda.textplots' utilise AI research results and create a Document Feature Matrix (DFM) for text mining analysis. The statistical and representational operations can only be performed on this data type (Benoit et al., 2018). The DFM matrix can be created from the frequency matrix (Kwartler, 2017).

Initially, the primary attributes of the notes were determined, namely the number of pages, rows, and characters. Next, the notes were cleaned by removing white space characters, numbers, filler words, punctuation, and control characters. After this, the occurrences of 263 terms that could be related to the given notes were examined. After completing these steps, a text file with quantitative data was created and converted into Microsoft Excel.

3.2. Text comparison

Text comparison is widely used in the social sciences. Various methods can help practitioners get information concerning the similarity of the texts. Some techniques can also quantify text differences. The

similarity between texts can be measured at different text levels, such as character strings, words (tokens), n-grams (n-unit character strings), and bags of words, but also between larger units of the documents, such as between text fragments and sentences (Sebők et al., 2021; Wang & Dong, 2020).

One measure commonly used to calculate lexical similarity is the Jaccard index, which is computed as follows (Niwattanakul et al., 2013):

$$Jaccard\ index\ (doc_{1,2}) = \frac{|doc_1 \cap doc_2|}{|doc_1 \cup doc_2|} \quad (1)$$

The Jaccard index helps determine the similarity between two documents. It is calculated by taking the number of words that match between the two documents and dividing it by the total number of words by the union of the total number of words in both documents. The resulting Jaccard similarity index gives the proportion of the words that are identical compared to the total number of words in the documents. This measure can benefit various applications, such as natural language processing and information retrieval (Wang & Dong, 2020).

4. EMPIRICAL RESULTS AND DISCUSSION

4.1. Descriptive statistics for the Term Frequency Matrix (TFM)

During the three years, 85,776 (28,592/year) documents were processed, containing 766,861 pages, 32,473,045 lines and 1,143,497,968 characters. The average size of a document was 8.94 pages, 378.58 lines and 13,331 characters in an average of years. Figure 1 shows the distribution of the examined notes per page size. Nearly 30% of the firms prepared notes of 1 and 5 pages. Most companies (41.95%) prepared notes between 6 and 10 pages. Only over 1% of the companies (1.11%) had notes larger than 30 pages. 79.63% of the companies examined prepared a simplified annual financial statement. Based on Figure 1, it can be concluded that most Hungarian enterprises did not prepare sufficiently detailed notes to the financial statements during the examined years. Specifically, 71.46% of the companies fell into the first two intervals, with an average number of pages of 6.08.

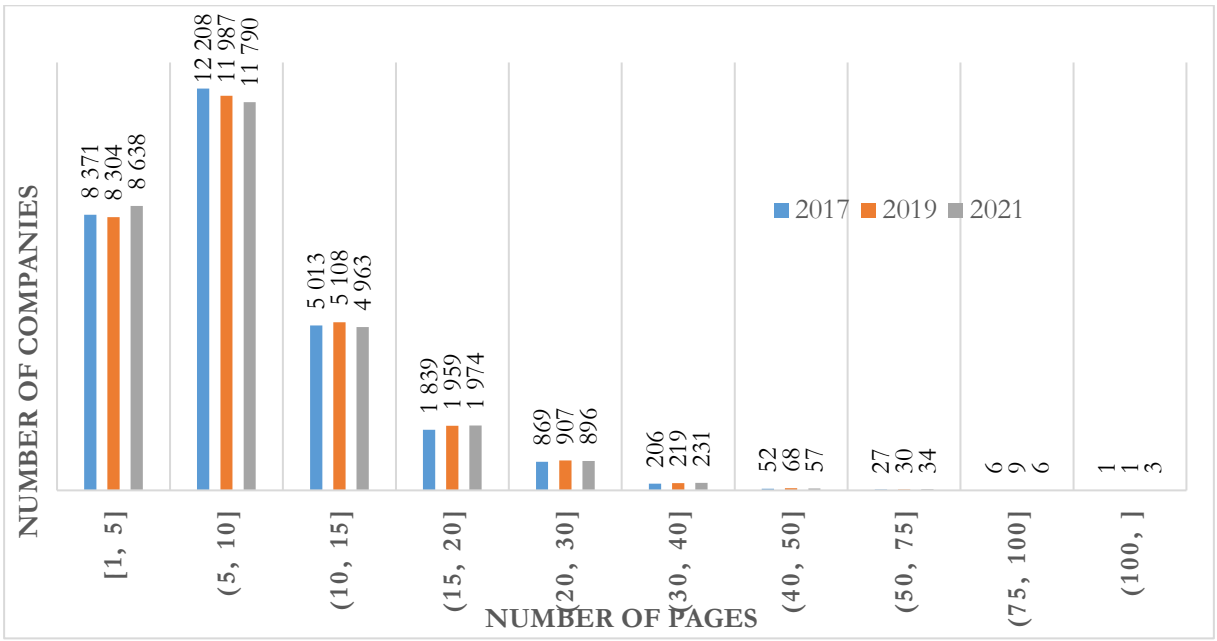


Figure 1. Distribution of companies based on the number of pages per year

Source: own editing

As the size of the company increases, it is expected that the length of the notes will also increase because of the need to explain more items. However, Figure 2 indicates no significant difference in the length of the notes in the first three size categories (average page numbers: 8.01, 7.88, 8.62). A noticeable difference is observed in the fourth category (average number of pages: 10.87), and a significant difference is evident only in the fifth category (average number of pages: 19.95). The coefficient of variation of the page numbers per size category hardly differs, with values ranging from 59.38% (category 2) to 65.37% (category 5). The Spearman correlation between the page number and company size is also very low, at 0.2058. This indicates that there is no relationship between page number and company size.

The terms being investigated were randomly chosen from the notes to financial statements of 100 companies of various sizes, each consisting of at least ten pages. All accounting and financial terms found in the selected notes were included in the analysis.

Figure 3 indicates that two terms, "tangibles" and "receivables", are present in over 90% of the notes. Three terms (inventories, depreciation and equity) occurred in more than 80% of the notes. Figure 3 also indicates that 121 out of the 262 terms (46%) appeared in less than 10% of the notes. 165 (63.36%) terms occur in less than 20% of the notes.

The study analysed the frequency of 262 accounting and financial terms in the notes, focusing more on the 67 most frequently used expressions (25.57%). In other words, only the expressions that occurred in at least 30% of the notes are analysed in more detail. Table 3 presents the statistical characteristics of the percentage of the notes using 262 and 67 terms over three years.

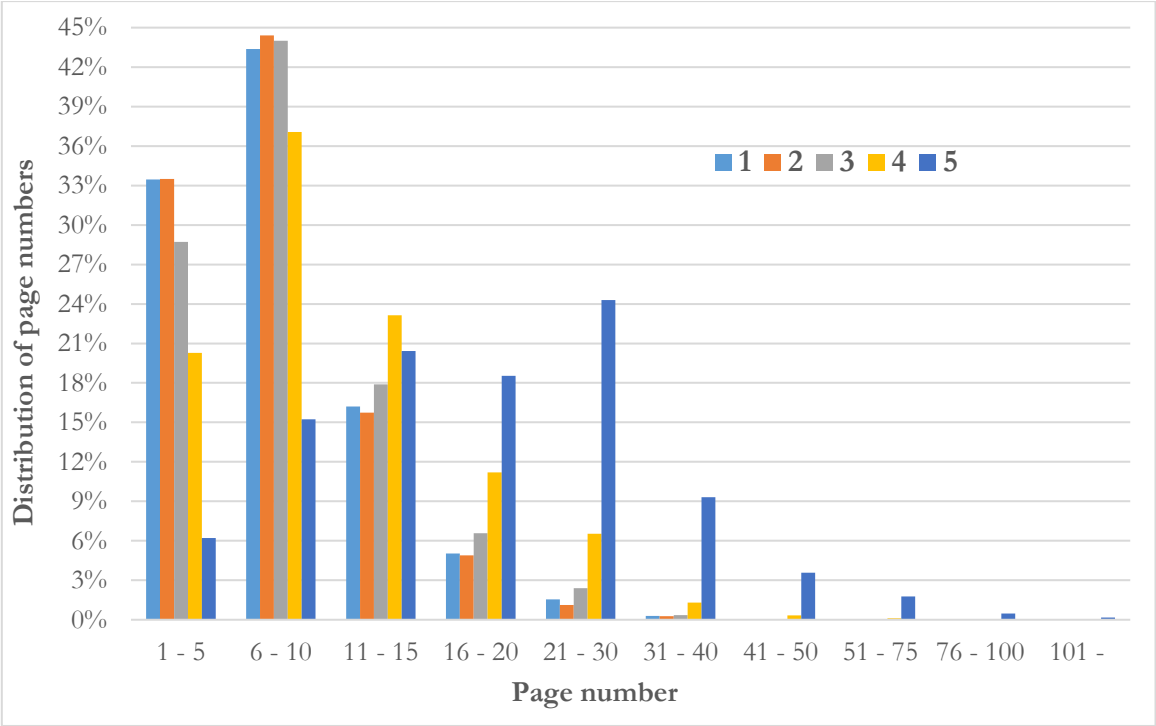


Figure 2. Distribution of companies by number of pages per company size

Source: own editing

The data in Table 3 shows the changes in the statistical characteristics of the notes. When focusing on 67 specific terms, compared to 262, the average frequency of term occurrences more than doubled, and the median value increased by almost five times. The coefficient of variant decreased significantly, from 107.31% to 29.64%, indicating a more uniform distribution. Moreover, the kurtosis and the skewness ratios

changed significantly, suggesting a shift from a distribution with a heavy tail to one with a thinner tail and a decrease in the number of outliers. The ratio of the interquartile range to the total range also indicates that the middle 50% represents a larger portion of the total range.

The low median and mean values in Table 3 indicate that many terms are absent from the notes to the financial statements. This suggests that the notes focus only on a few key areas. Upon analysing data from all the companies over three years, it was discovered that only 38 out of the 262 terms appeared at least once on average in the Term Frequency Matrix.

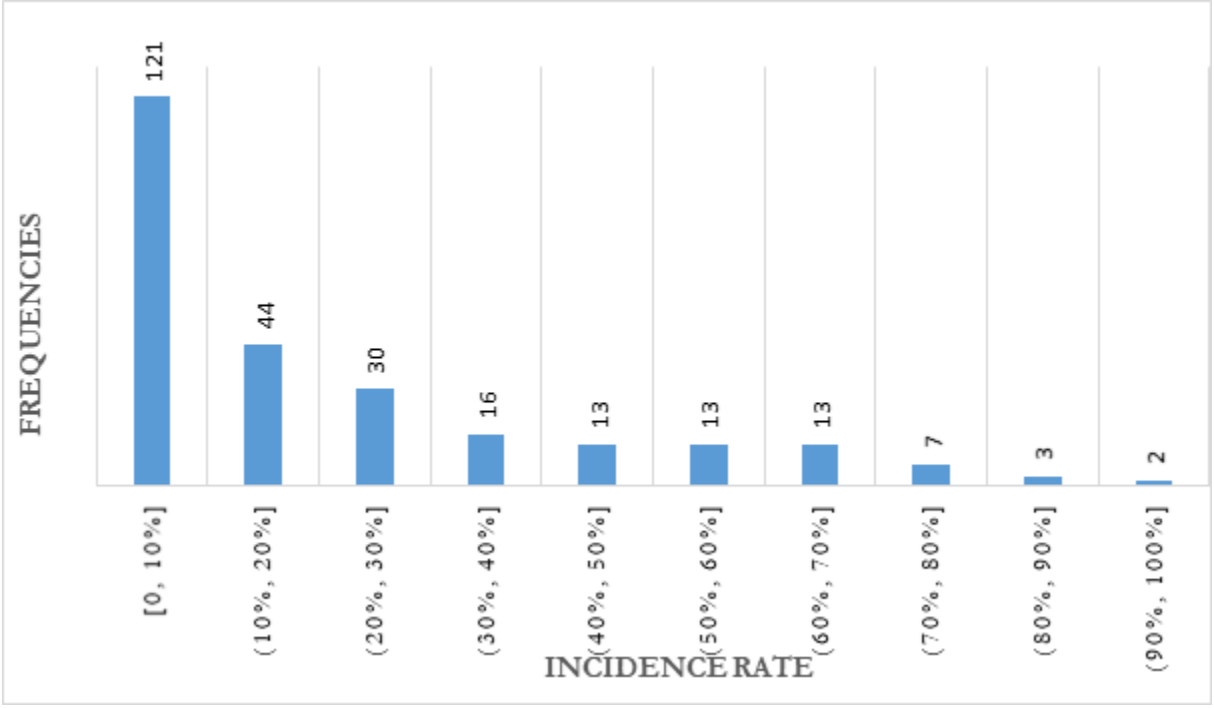


Figure 3. The number of terms by proportion to their average occurrence

Source: own editing

Table 3

Statistical characteristics of the proportion of the notes that used examined terms

Indicator names	Statistical indicators of term frequencies	
Number of terms	**262**	**67**
Minimum	0.00%	30.42%
Quartile 1	4.07%	40.35%
Median	11.59%	52.14%
Quartile 3	30.99%	66.39%
Maximum	93.39%	93.39%
Mean	21.11%	54.80%
Standard deviation	22.65%	16.24%
Coefficient of variant	107.31%	29.64%
Skewness	1.2340	0.4595
Kurtosis	0.5741	-0.6956
IQR (interquartile range)	26.91%	26.04%
Total range	93.39%	62.97%
IQR/Total range	28.81%	41.36%

Source: own editing

Figure 4 shows the term cloud of the 67 most frequently used terms, with font size indicating the term frequency. The expression frequencies presented in Figure 4 align with those described in Figure 3. The three most frequently used terms are 'tangibles', 'receivables' and 'equity', followed by 'inventories', 'intangibles', 'profit after tax', and 'depreciation'. From this, it can be concluded that the companies primarily focus on their assets and financial results in their notes.

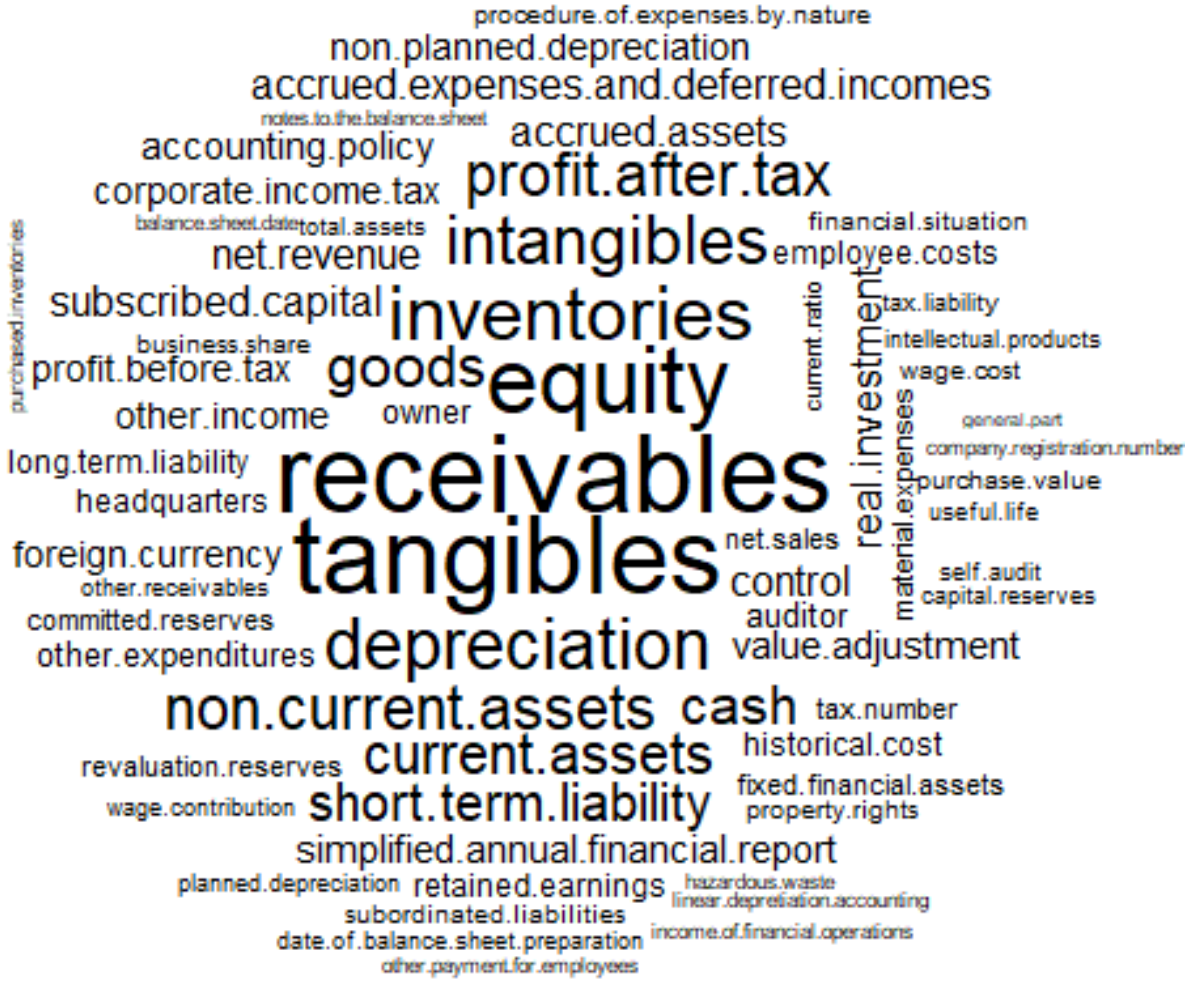


Figure 4. Term cloud of the 67 focused terms

Note: Dots separate the expressions because they are used as variable names in the R statistical system, which does not allow spaces in variable names.

Source: own editing

Table 4 includes terms that appear in at least two-thirds of the notes. Out of 262 terms, only 16 (6.08%) are listed in the table, about 23.88% of the total 67 terms. The terms are sorted by their frequency of occurrence, indicating how many proportions of notes they appear in at least once. On average, 14 of the 16 terms occur at least once in the examined notes. Of the 67 terms, 40 (59.70%) appear at least once in the notes. However, only 13 expressions occur at least twice. The previous data draws attention to the fact that the notes may be significantly incomplete.

Table 4

The proportion of notes containing the most frequent terms

Num.	Terms	Proportion of notes containing terms	Average occurrence of terms
1	tangibles	93.39%	5.98
2	receivables	92.71%	6.16
3	inventories	86.63%	3.97
4	depreciation	83.08%	3.62
5	equity	82.70%	5.22
6	simplified annual financial report	79.63%	1.63
7	non-current assets	79.35%	3.06
8	current assets	75.75%	2.77
9	intangibles	74.02%	3.57
10	profit after tax	72.33%	3.04
11	goods	71.49%	2.80
12	headquarters	70.24%	1.09
13	accounting policy	69.48%	1.60
14	accrued expenses and deferred incomes	69.37%	1.78
15	procedure of expenses by nature	69.29%	0.77
16	tax number	68.59%	0.93

Note: The average occurrence of terms equals the total occurrence of terms divided by the number of notes.
Source: own editing

Figure 5 indicates that the data does not follow normal distribution, further confirmed by the Kolmogorov-Smirnov test (p-values are less than 0.001 for the three years). The figure also illustrates minimal variations between the different years, which are not considered statistically significant, according to the Kruskal-Wallis test (p-value = 0.9893). The black line in the figure depicts the average occurrences of terms over three years. Out of 67 terms, over 65 are found in only one note, and over 60 are found in 251 (0.88%) ones. On average, over three years, over 50 terms are found in 4,659 companies (16.29%). In almost 41% of the notes, no more than 33 of the terms analysed are present. The Spearman correlation between the term occurrence and company size is also very low, at 0.1691. This indicates that there is no relationship between term occurrences and company size. Based on the findings, it is likely that many companies do not adequately prepare detailed notes.

4.2. Factor impacts on notes to financial statements

A Spearman correlation analysis was conducted using the dimensions of the notes and the available factors (Table 5). The analysis revealed that the different size characteristics of the notes are closely related, while these values only show a weak relationship with corporate size. So, notes with more pages have more lines and characters. Additionally, there is a weak negative correlation between the year and the number of rows.

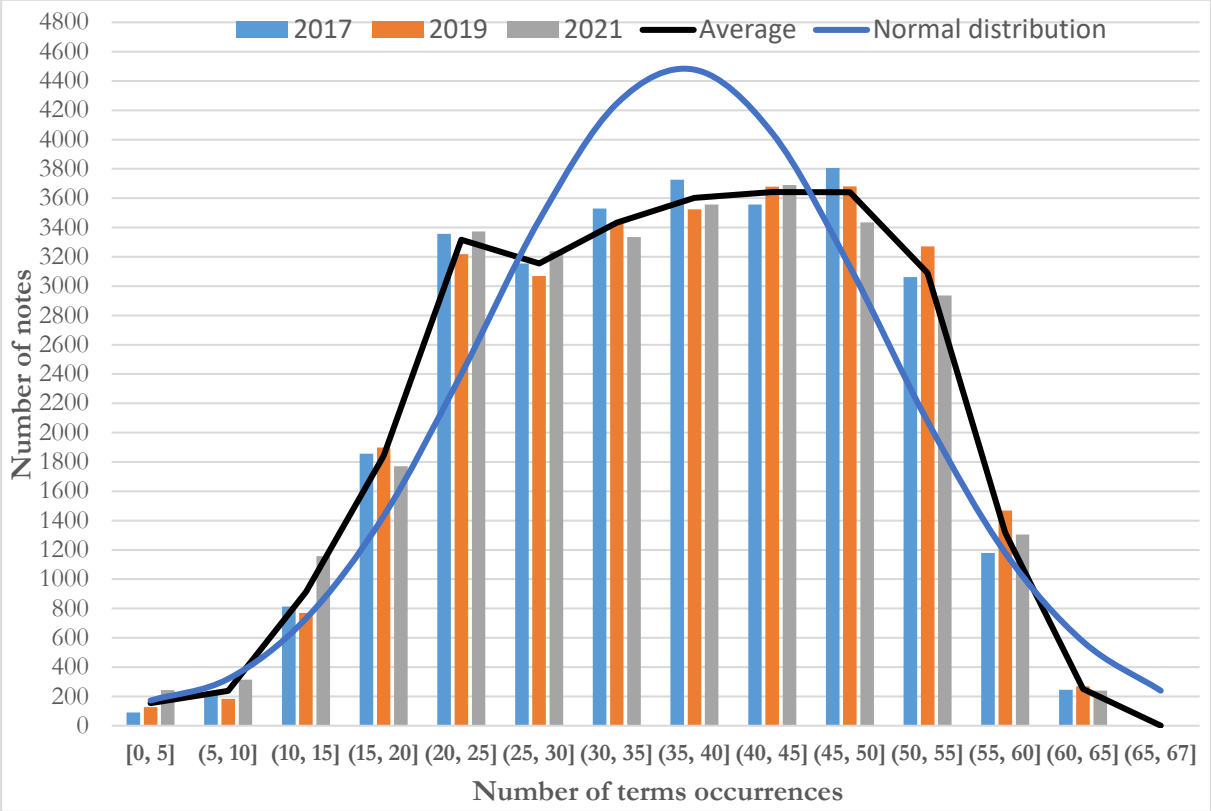


Figure 5. Distribution of notes based on term occurrence per year (Number of terms = 67, Number of notes per year = 28 592)

Source: own editing

Table 5

Spearman correlation coefficients

Variable names	Year	Sector	Company size	Pages	Rows	Characters without white spaces	Characters without numbers	Characters without punctuations
Year	1.0000	0.0000	0.0000	-0.0035	-0.2422	0.0057	-0.0088	-0.0051
Sector	0.0000	1.0000	-0.1876	-0.0379	-0.0325	-0.0332	-0.0320	-0.0304
Company size	0.0000	-0.1876	1.0000	0.2058	0.1989	0.2103	0.2050	0.2000
Pages	-0.0035	-0.0379	0.2058	1.0000	0.9188	0.9046	0.8980	0.8886
Rows	-0.2422	-0.0325	0.1989	0.9188	1.0000	0.9148	0.9152	0.9061
Characters without white spaces	0.0057	-0.0332	0.2103	0.9046	0.9148	1.0000	0.9985	0.9956
Characters without numbers	-0.0088	-0.0320	0.2050	0.8980	0.9152	0.9985	1.0000	0.9981
Characters without punctuations	-0.0051	-0.0304	0.2000	0.8886	0.9061	0.9956	0.9981	1.0000

Source: own editing

According to the Kolmogorov-Smirnov test, the tested variables do not show a normal distribution. Because of these results, the Kruskal-Wallis test was used to assess the impact of factors (year, sector, company size) on the variables tested. A multivariate Kruskal-Wallis test was performed incorporating all five variables, and the effects per variable were also determined (Table 6). Based on the table, the multivariate Kruska-Wallis test indicates statistically significant differences in the notes based on the year, sector and company size. Additionally, differences based on all factors have a significance level of at least 1% for the five variables. In summary, it can be inferred that the factors examined statistically impact the size of the notes to financial statements.

Table 6

Results of the Kruskal-Wallis test considering the overall characteristics of the notes

Tests	Year	Sector	Company size
Multivariate Kruskal-Wallis test	***	***	***
Number of pages	**	***	***
Number of rows	***	***	***
Character number without white space	***	***	***
Character number without white space and numbers	***	***	***
Character number without white space, numbers and punctuation	***	***	***

Notes: *** significance level <= 0,1%
** significance level <= 1%
'- no significant difference

Source: own editing

4.3. Statistical comparison of term frequencies

The Kruskal-Wallis test assesses if there are statistically significant variances between the medians of more than two independent groups. If the result of the Kruskal-Wallis test is statistically significant, then Dunn's Test is appropriate for determining which groups are different. The Kruskal-Wallis test was used to determine if there was a difference in the individual and total occurrences of the term 67, considering the year, size, and sector factors.

First, it was tested whether the frequency of word occurrence differs from year to year. Kruskal-Wallis revealed a statistically significant difference between years at less than 0.1%. Subsequent Dunn test confirmed that the difference between each pair of years was also statistically significant at a level of less than 0.1%.

As a next step, the notes were compared annually according to the five size categories. The Kruskal-Wallis test indicated a significant difference, at least at the 0.1% level every year. This suggests that the frequency of the terms of notes depends on the company size overall. The Dunn test, comparing term frequencies based on company size, also revealed a significant difference, at least at the 0.1% level, except for pairs 1-2. Notably, the difference between pairs 1-2 was not statistically significant in either year. The results confirm that the examined companies updated their notes from the previous year and created them with different term frequencies.

Figure 6 illustrates the average occurrence of 67 terms at least once in grouping notes by company size categories. The figure indicates no significant difference in the occurrence of the terms between the first two categories, as supported by the Dunn test. The difference is only 0.37% in favour of the 2nd category.

However, there is an increased difference in occurrence between size categories 4 and 5. The results suggest two possible conclusions. Firstly, larger companies may have more items to explain on their balance sheets and income statements. Secondly, it could be inferred that more attention is given to the preparation of the notes. The latter seems more plausible since larger companies have greater administrative resources for compiling notes.

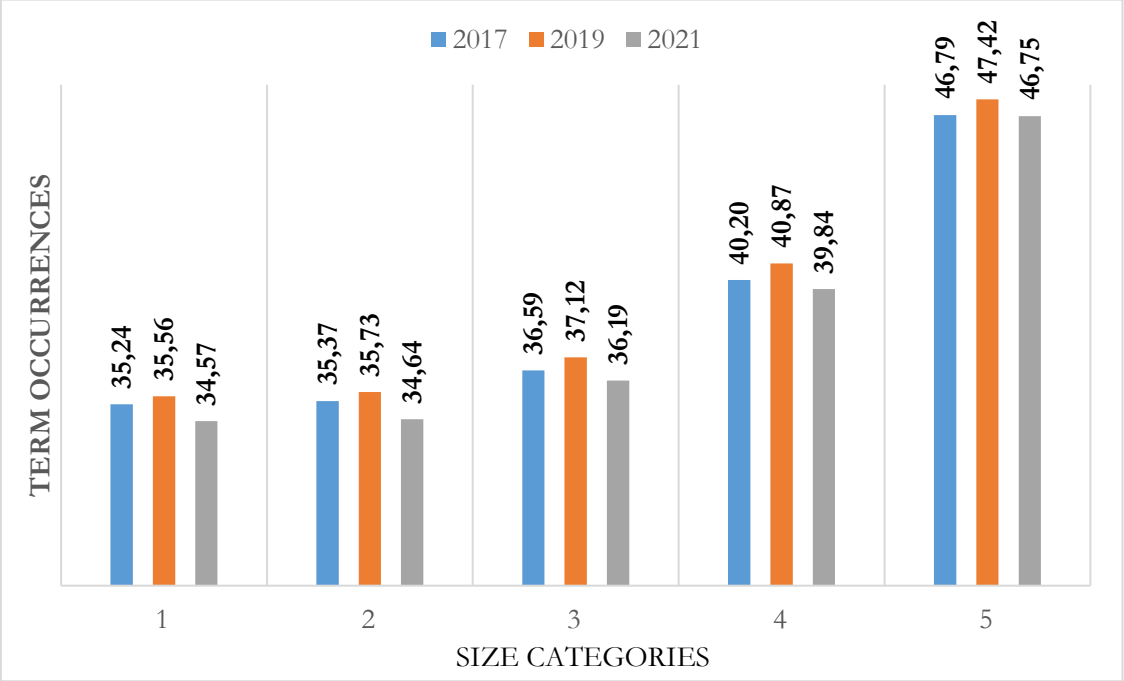


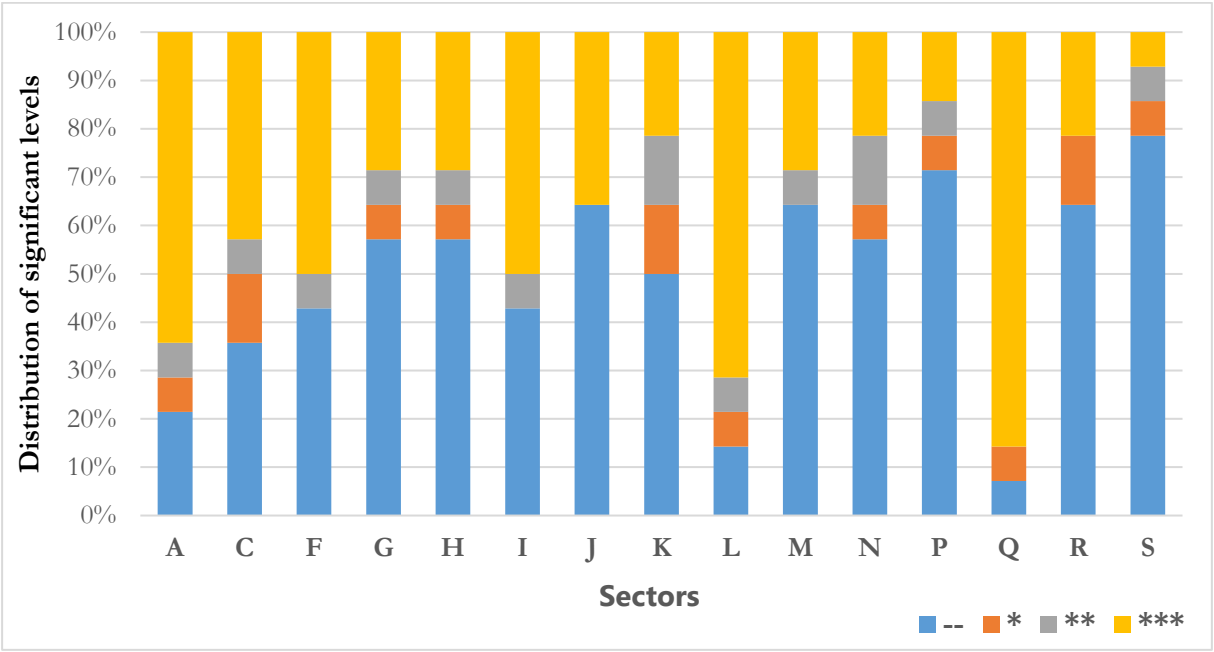
Figure 6. Average single occurrence of terms in notes grouping by company size (Number of terms examined = 67)

Source: own editing

In Figure 7, the results of the sector comparisons applying the term frequency matrix are displayed. The comparison per sector was conducted using the Dunn test. The figure indicates that the Q (Human Health and Social Work Activities) sector is most different from other sectors, where 92.86% of pair-wise comparisons show a significant difference of at least 5%. Figure 8 illustrates why the Q sector differs significantly: it has the lowest average occurrences of terms in the notes. This is followed by the L (Real estate activities - 85.71%) and A (Agriculture, Forestry, and Fishing—78.57%) sectors. The higher term occurrence causes a significant difference in sectors L and A, opposite to the Q sector (Figure 8). On the other hand, the S (Other service activities - 21.43%) sector is the least different from other sectors. There are differences among the terms used by sectors, and the discrepancies are very different.

4.4. Similarity analysis

The Jaccard Index was used to conduct the similarity test. During the analysis, all notes were compared to each other, resulting in a large matrix (28.592 * 28.592) yearly. The Jaccard index values were then placed in frequency intervals. The frequency values of the intervals of the various features (year, sector, corporate size) were compared by chi-square test.



A	C	F	G	H	I	J	K	L	M	N	P	Q	R	S
78.57%	64.29%	57.14%	42.86%	42.86%	57.14%	35.71%	50.00%	85.71%	35.71%	42.86%	28.57%	92.86%	35.71%	21.43%

Figure 7. Average single occurrence of terms in notes grouping by industry and the proportion of statistically significantly different notes (Number of terms examined = 67)

Source: own editing

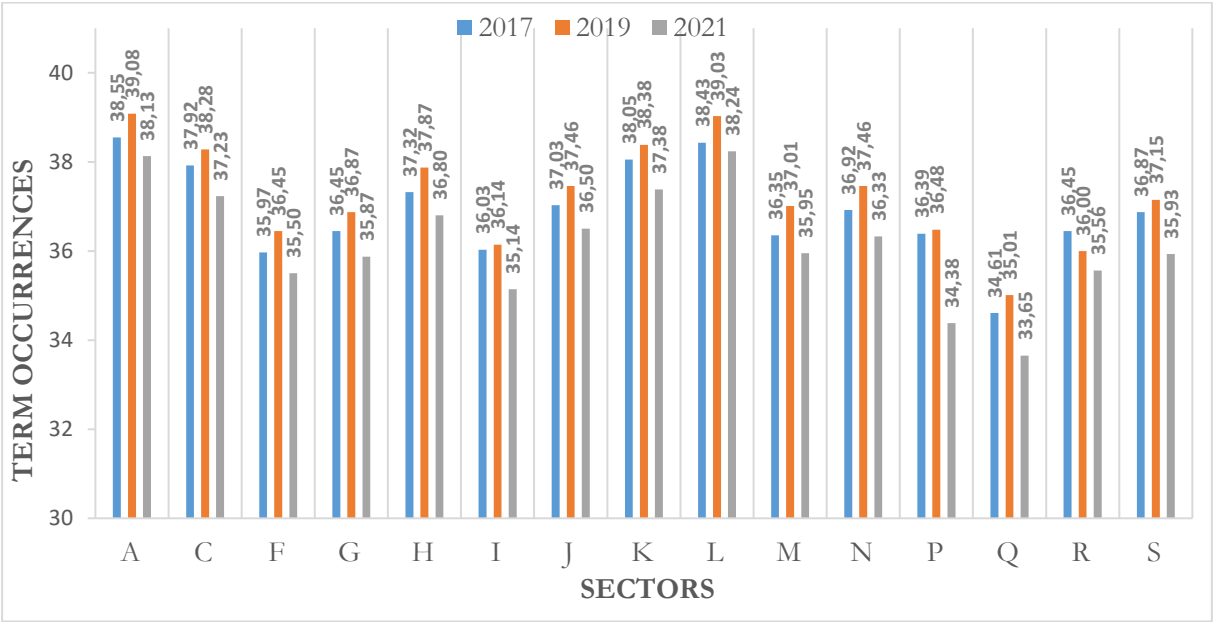


Figure 8. Average single occurrence of terms in notes, grouped by sector and year

Source: own editing

Figure 9 displays the distribution of yearly comparisons of the Jaccard index. The figure indicates that there is not much variation in the distribution of the index. However, the Chi-square test revealed significant differences between the comparisons. The comparison between 2017 and 2021 showed the largest test value, while the comparison between 2017 and 2019 showed the smallest one. The average Jaccard similarity indices with any yearly comparison do not exceed 0.3. This suggests that the notes have varied over the years and have not been duplicated.

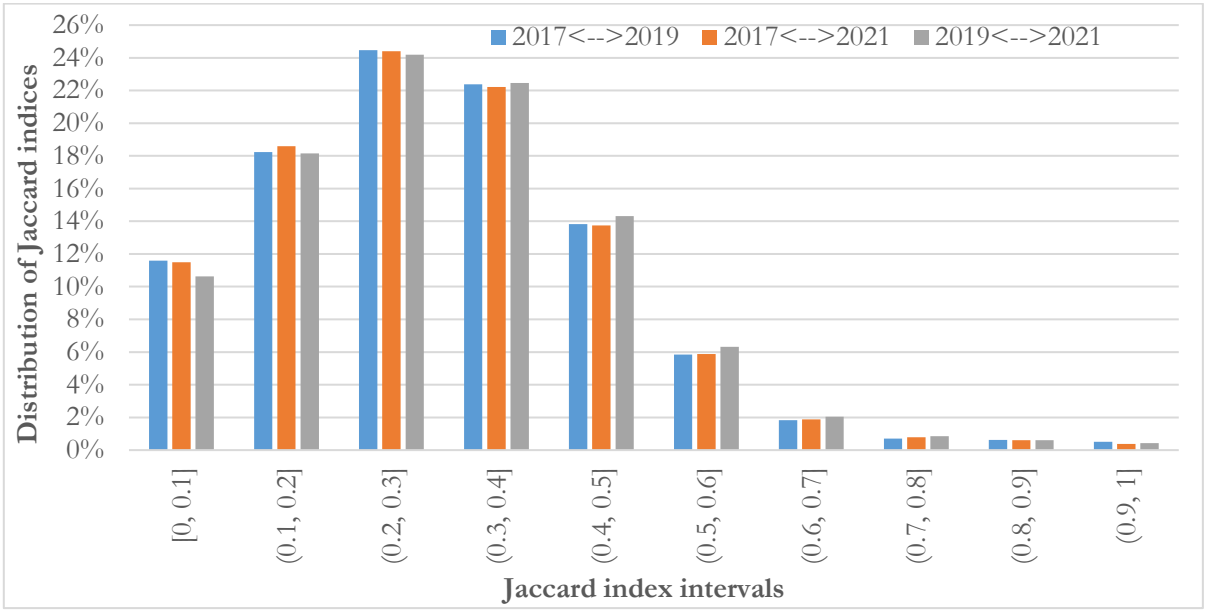


Figure 9. Distribution of the Jaccard index per year

Source: own editing

The next step was determining the Jaccard index distribution by corporate size comparison (Figure 10). The figure shows significant differences between the comparisons made. The average Jaccard index values per comparison of company sizes range from 0.399 to 0.534, and the standard deviation of the values is not too significant. The coefficients of the variant are close to 30%, with the highest value being only 37.57%, which is reasonable considering the large number of elements involved. All Jaccard indices based on corporate size comparison differ significantly, considering the chi-square test, at least on a 0.1% level. The biggest similarity is between size categories 4 and 5, where 31.23% of the notes exceed the similarity index of 0.7; the smallest is between size categories 1 and 2, where this ratio is 9.26. Based on the results, it can be inferred that the company's size impacts the contents of the notes.

The third comparison was carried out for each sector, resulting in 105 comparisons. Due to the large number of elements, only a few characteristics can be represented, as illustrated by the box-plot diagram in Figure 11. The box-plot diagram displays the dataset by quartiles, with each quarter containing the same number of elements. The figure shows that the range of the first quartile is the largest and the third is the smallest. It also shows that the difference between Jaccard indexes of comparisons is not too high (max - min = 0.065; 14% compared to the minimum). Additionally, chi-square tests reveal significant differences between comparisons. This suggests that similarity indexes vary among sectors, indicating differences in the content of the notes across industries.

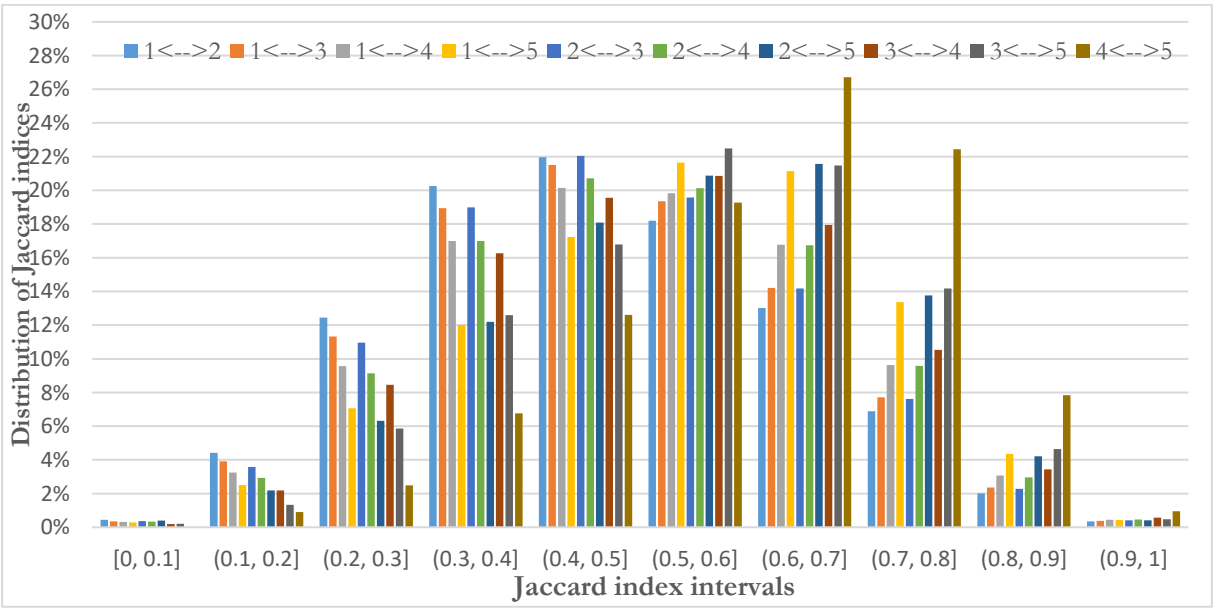


Figure 10. Distribution of the Jaccard index per company size

Source: own editing

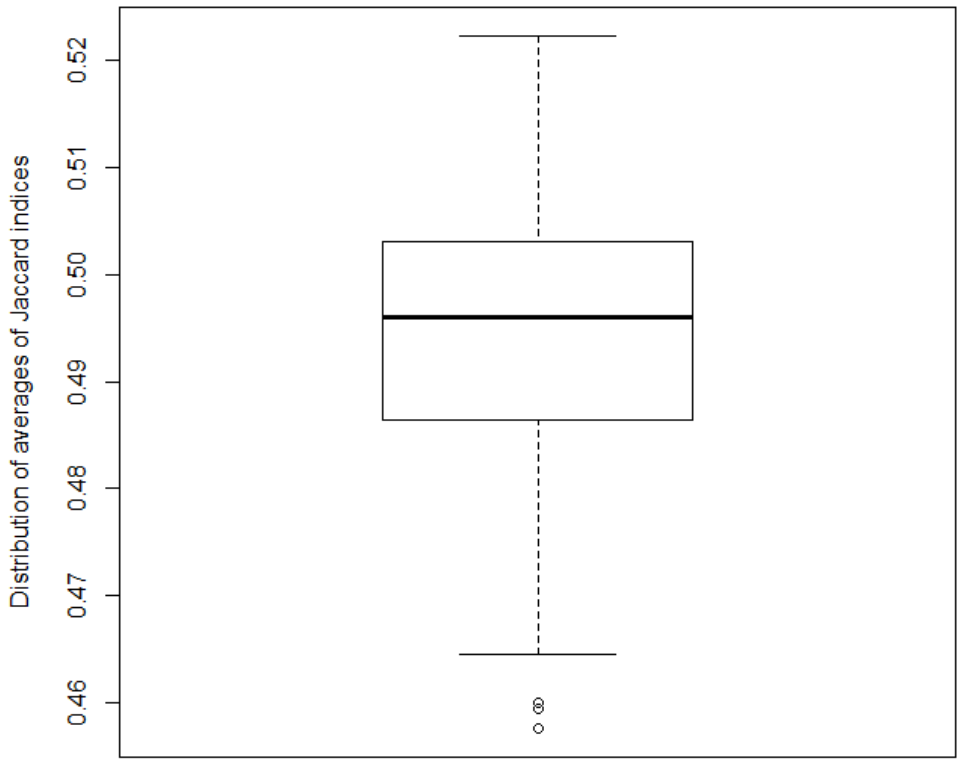


Figure 10. Distribution of the Jaccard index per sector using a box-plot diagram

Source: own editing

5. CONCLUSION

This study analysed the Hungarian financial statement notes for 2017, 2019, and 2021. The study is unique because analysing so many companies' notes to financial statements has not yet been done in Hungary before. The research resources limited the bi-yearly analysis of the notes to the financial statements. However, this provides an appropriate opportunity to examine the changes. The 2023 notes to the financial statements were not yet available at the time of the analyses or when the study was prepared.

Text mining was utilised as the primary method for analysing the notes to the financial statements. The text mining method, based on the artificial intelligence available in the R statistical system, was highly effective, even for such a huge database.

The analyses indicate that the notes use only a limited number of the selected accounting and financial terms. This finding is consistent with analyses based on a much smaller sample from specific sectors at an early stage. The tests also revealed that the notes published by companies are often incomplete and lack essential information for stakeholders.

Starting in 2025, companies will be required to prepare ESG (Environmental, Social and Governance) reports under the regulation of the European Union. These reports will give stakeholders a broader understanding of the company (Dathe et al., 2024). However, it is important to note that ESG reports cannot replace the notes to the financial statements. The regulatory authorities should ensure that the notes and other parts of the statements are presented in a specific structure, facilitating easier analysis and control. Currently, there are no checks on these documents, and in many cases, other documents unrelated to notes, such as General Assembly or income allocation decisions, have been mistakenly uploaded.

The original database included different documents instead of notes, such as general meeting decisions or the results divisions. Some notes only had a title consisting of just one line. These problematic notes made up close to 10% of the entire database. No one verifies financial statements published in Hungary.

Analysing the term frequency matrix for the 67 most common terms has revealed no significant difference between the years. However, considerable differences have been caused by size categories and sectors. It should also be noted that the last two categories have influenced the differences in size categories.

It has been confirmed that the notes are statistically significant using Jaccard similarity analysis, considering the year, corporate size, and sector. However, it is important to note that the results of the statistical tests depend significantly on the sample size, and this analysis had very high elements considering the similarity analysis.

The study can help convince regulatory authorities to change the current system. The change should affect the entire financial reporting system, as the situation is inadequate for other parts of the report (balance sheets, income statements). The balance sheets and income statements are recorded digitally, but the control does not work for them.

Presentation of the method used can encourage further research on the notes of the financial statements and ESG reports.

The analyses did not cover all the potential examinations because of limitations in the size of the paper. In the future, consider using methods that can provide a deeper insight into the notes' features, such as machine learning, deep learning, various Bayes statistics, and techniques associated with the Python programming language for accounting and finance.

6. LIMITATIONS OF RESEARCH

Current research has three limitation factors. The first is that according to the criteria provided by us for Opten Informatikai Ltd., they have compiled the research database, which we could not wholly verify. The second limit was that the available financial research sources allowed the purchase of the database only

every two years. Third, we can get the database by delaying because we depend on Opten Kft. Therefore, the research does not yet include an analysis of 2023. We also plan to purchase and analyse the notes of 2023.

ACKNOWLEDGEMENT

This paper supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the University of Debrecen Program for Scientific Publication.

REFERENCES

Abernathy, J.L., Guo, F., Kubick, T.R., & Masli, A. (2018). Financial Statement Footnote Readability and Corporate Audit Outcomes (27 August 2018). *Auditing: A Journal of Practice & Theory, Forthcoming*, Available at SSRN: https://ssrn.com/abstract=3239625 https://doi.org/10.2308/ajpt-52243

Amani, F.A., & Fadlalla, A.M. (2017). Data mining applications in accounting: A review of the literature and organising framework. *International Journal of Accounting Information Systems*, 24(2017), 32-58. http://dx.doi.org/10.1016/j.accinf.2016.12.004

Aymen, A., Sourour, B.S., & Badreddine, M. (2018). The effect of annual report readability on financial analysts' behaviour. *Journal of Economics, Finance and Accounting* (JEFA), 5(1), 26-37. http://doi.org/10.17261/Pressacademia.2018.782

Bai, X., Dong, Y., & Hu, N. (2019). Financial report readability and stock return synchronicity. *Applied Economics*, 51(4), 346-363. https://doi.org/10.1080/00036846.2018.1495824

Barnett, A., & Leoffler, K. (1979). Readability of Accounting and Auditing Messages. *International Journal of Business Communication*, 16, 49-59. https://doi.org/10.1177/002194367901600305

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software***.** 3(30), 774-777. http://dx.doi.org/10.21105/joss.00774

Bíró, F. P., Erdey, L., Gáll, J., Márkus Á. (2019): The Effect of Governance on Foreign Direct Investment in Latin America – Issues of Model Selection. *Global Economy Journal*, 19(1), 97-116. https://doi.org/10.1142/S2194565919500064

Bohusova, H., Svoboda, P., Veverkova, A. (2022). Impact of New Lease Reporting on Retailing and Wholesale Companies. *Montenegrin Journal of Economics*, *18*(3), 89-98. DOI: 10.14254/1800-5845/2022.18-3.7

Chan, S.W., & Franklin, J. (2011). A text-based decision support system for financial sequence prediction. *Decision Support Systems*, 52(1), 189-198. https://doi.org/10.1016/j.dss.2011.07.003

Dathe, T., Helmold, M., Dathe, R., & Dathe, I. (2024). Implementing Environmental, Social and Governance (ESG) Principles for Sustainable Businesses. A practical guide in sustainability management. Springer Nature Switzerland. ISBN 978-3-031-52734-0

Feinerer. I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*. 25(5), 1-53**.** ISSN:1548-7660

Fenyves, V., Böcskei, E., Bács, Z., Zéman, Z., & Tarnóczi, T. (2019). Analysis of the Notes to the Financial Statement Related to Balance Sheet in Case of Hungarian Information-Technology Service Companies, *Scientific Annals of Economics and Business* 66 (1), 27-39. https://doi.org/10.2478/saeb-2019-0001

Filyó, J. (2014). A kiegészítő melléklet ellenőrzésének tapasztalatai. [Experiences of checking the notes to financial statements.] *Számvitel, adó, könyvvizsgálat [Accounting, taxation, auditing],* 56(6), 285-286. ISSN 1419-6956.

Földvári, P. & Erdey, L. (2009): Do Purchasing Power and Interest Rate Parities Hold for the EUR/HUF exchange rate? A time-series analysis. *Acta Oeconomica*, 59(3), 289-306. https://doi.org/10.1556/aoecon.59.2009.3.2

Gandía, J.L., & Huguet, D. (2021). Textual analysis and sentiment analysis in accounting: Análisis textual y del sentimiento en contabilidad. *Revista de Contabilidad-Spanish Accounting Review*, 24(2), 168-183. https://dx.doi.org/10.6018/rcsar.386541

Gupta, R., & Gill, N.S. (2012). A solution for preventing fraudulent financial reporting using descriptive data mining techniques. *International Journal of Computer Applications*, 58(1), 22-28. https://doi.org/ 10.5120/9247-3411

Jofre, M., & Gerlach, R. (2018). Fighting accounting fraud through forensic data analytics. *SSRN Electronic Journal*. January 2018, 1-39. http://dx.doi.org/10.2139/ssrn.3176288

Kántor, B. (2016). A kiegészítő melléklet. [Notes to financial statements] *Számviteli Tanácsadó [Accounting advisor]*, 8(3), 2-11. ISSN 2060-4076

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171-185. https://doi.org/10.1016/j.irfa.2014.02.006

Kerezsi, D. (2017). A kiegészítő melléklet szerepe a piaci szereplők tájékoztatásában. [The role of the notes to financial statements in informing market participants] *International Journal of Engineering and Management Sciences / Műszaki és Menedzsment Tudományi Közlemények, 2*(4), 202-212. https://doi.org/10.21791/IJEMS.2017.4.17

Kerezsi, D., Béresné Mártha, B., & Sütő, D. (2019). Sector analysis of the Notes in Northern Great Plain region's enterprises. *Controller Info*, 7(3), 47-50. https://doi.org/10.24387/CI.2019.3.10

Kerezsi, D. (2020). To what extent does the information disclosure of sports and ICT companies comply with the legal requirements? *Annals of the University of Oradea Economic Science*, 29(1), 240-251.

Kerezsi, D. (2021). The role of the bank loan related information of the notes in entrepreneurial decision-making - evidence from Hungarian entreprises. *Network Intelligence Studies*, 9(18), 95-105.

Kwartler, T. (2017). Text Mining in Practice with R. John Wiley & Sons Ltd. ISBN 9781119282099

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230. https://doi.org/10.1111/1475-679X.12123

Lepadatu, G.V., & Pirnau, M. (2009). Transparency in financial statements (IAS/IFRS). *European Research Studies Journal*, 12(1), 101-108. https://doi.org/10.35808/ersj/212

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard Coefficient for Keywords Similarity. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol I, IMECS 2013, March 13-15, Hong Kong. ISSN: 2078-0966

Osadchy, E.A., Akhmetshin, E.M., Amirova, E.F., Bochkareva, T.N., Gazizyanova, Yu.Yu., & Yumashev, A.V. (2018). Financial statements of a company as an information base for decision-making in a transforming economy. *European Research Studies Journal*, 21(2), 339-350. https://doi.org/10.35808/ersj/1006

Pakšiová, R., & Oriskóová, D. (2020). Capital maintenance evolution using outputs from accounting system. *Scientific Annals of Economics and Business*, 67(3), 311-331. https://doi.org/10.47743/saeb-2020-0017

Shakatreh, M., Abu Orabi, M.M., Al Abbadi, A.F.A. (2023). Impact of Cloud Computing on Quality of Financial Reports With Jordanian Commercial Banks. *Montenegrin Journal of Economics*, *19*(2), 167-178. https://doi.org/10.14254/1800-5845/2023.19-2.14

Sebők, M. (ed.) (2016). Kvantitatív szövegelemzés és szövegbányászat a politikatudományban" (Quantitative text analysis and text mining in political science), L'Harmattan Kiadó.

Senave, E., Jans, M.J., & Srivastava, R.P. (2023). The application of text mining in accounting. *International Journal of Accounting Information Systems*. 50(2023), 100624. https://doi.org/10.1016/j.accinf.2023.100624

Thalassinos, I.E., & Liapis, K. (2014). Segmental financial reporting and the internationalisation of the banking sector. Chapter book in, Risk Management: Strategies for Economic Development and Challenges in the Financial System, (eds), D. Milos Sprcic, Nova Publishers, 221-255, ISBN 978-1633214965

Tóthné Szabó, E. (2010). A kiegészítő melléklet szerepe a "megbízható és valós kép" kialakításában. [The role of the notes to financial statements in creating a "reliable and real image"]. Számvitel Adó Könyvvizsgálat [Accounting, taxation, auditing]: SZAKMA, 52(4), 180-186.

Wang, J., & Dong, Y. (2020). Measurement of Text Similarity: A Survey. *Information*, 11(9), 421–437. https://doi.org/10.3390/info11090421

Yadav, A.K.S., & Sora, M. (2021). Fraud detection in financial statements using text mining methods: A review. In IOP conference series: Materials science and engineering. 1020(1), 1:19, (012012). IOP Publishing. https://doi.org/10.1088/1757-899X/1020/1/012012