Soft Computing and Machine Intelligence

Soft Computing and Machine Intelligence Journal, Vol (1), Issue (2), 2021

Analyzing the Performance of User Generated Contents in B2B Firms Based on Big Data and Machine Learning

Zeinab Shahbazi [a] **and Yung-Cheol Byun** [b,*]

[a,b] Department of Computer Engineering, Jeju National University, Jeju-si 63243, Jeju Special Self-Governing Province,Korea; zeinab.sh@jejunu.ac.kr

* Corresponding author: ycb@jejunu.ac.kr

Abstract: Social media platforms act as a significant role in human life in recent decades. Marketing scholars show interest in the field of big data based on user-generated content from social media platforms. However, maximum user-generated content is conducted in terms of business-to-consumer (B2C) context to improve the knowledge differences in business to business (B2B) area. The dataset used in the proposed system collects from the Twitter platform. The extracted information is related to eight years of stock data related to 407 companies. Similarly, machine learning techniques are applied to predict data performance. The result of machine learning is converted to the monthly panel dataset. Based on the analysis results, user-generated contents have a considerable impact on companies, showing the differences between B2B and B2C firms. The generated results show that B2C performance is higher and more reliable than B2B. In this process, the consumer's positive response does not affect the stock data performance.

Keywords: Big Data, Machine Learning, Sentiment Analysis, User-generated Content

1. Introduction

One of the important digital age sides is an unstructured dataset. Tweets, blogs, Facebook shared posts, news, and reviews on various topics or websites are full of user behaviors, actions, and their vision which fetch unexampled potential time to researchers and companies [1,2]. User-generated content based on big data is a popular topic among marketing scholars. Based on the development in the research area, most of the researchers try to apply new methods on big data to improve the understanding of users' behavior and activities on social media. One of the surveys in the USA describes 84% of industrial firms using big data analysis to get better accuracy on decision making. Big data analysis, i.e., the use of big data and related methods, are presented in most of the firms' values based on reducing costs and making new ways for alteration and disordering [3,4]. The main reason is the process of data is to make the business knowledge [5], i.e., understanding the business process and getting some related information is the best option in decision making in firms' area [6,7]. Similarly, the analysis process can use for optimization process in business to thoroughly understand the customers' interests [8,9]. This process can extract the hidden patterns of customer's behavior too. Most of the user-generated content is used in business- to- consumer (B2C) context. To cover the gap in B2B research, marketing scholars are encouraged to use social media contents advantages and applying new ideas related to artificial intelligence and big data. The primary issue for this process is data collection, data pre-processing and data analysis of the user-generated contents [10]. This paper inquires the combination of big data and machine learning in B2B settings. Similarly, the relationship between user-generated content and stock performance companies is accessible publicly [11]. Data

collection is based on java programming for crawling social media information. Eighty-four million tweets were extracted from 20 million user accounts. Similarly, 407 companies, eight years of stock data, were also extracted. Figure 1 presents the process of data collecting and data analysis. The process is dividing into five sections: data collection, data pre-processing, data visualization, data aggregation, and data analysis. The data collection step, identify the dataset and the primary consideration is on data issues which the input of this process is an unstructured raw dataset. The second step is data pre-processing which discovers spam and tokenizes the dataset, and similarly makes a structure for the dataset. The third step is data visualization which the main focus is to visualize the data set and apply the LDA topic modeling technique for extracting the hidden information out of the document [12,13]. The fourth step is data aggregation which integrates the dataset using an SQL server. Finally, the last step is data analysis which finalizes the whole analysis on the dataset and shows the result of topic modeling. The main contribution of this paper is divided into five sections. Section 2 representing the related studies, section three represents the proposed method and result, and finally, we conclude the paper in section five.

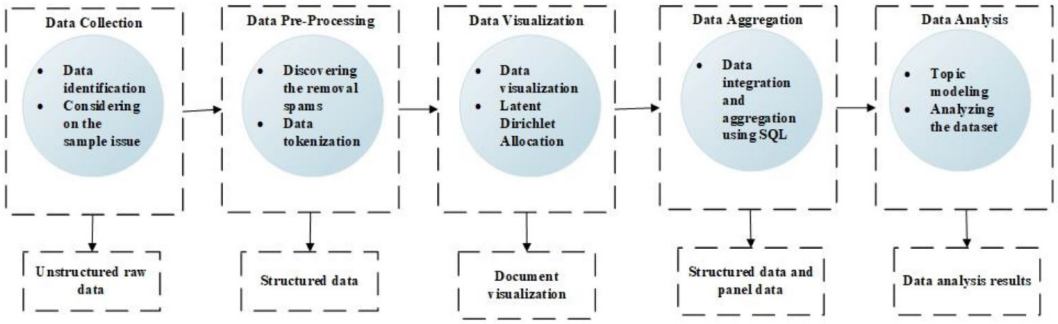

Figure 1. Data Analysis Process Using Big Data and Machine Learning

2. literature review

User-generated content is designed based on general consumers in professional marketing [14]. Based on the increase of users in social media websites and the development of the internet, a large amount of user-generated content, e.g., user-product feed-backs, Facebook posts, tweets, shared information , and videos on YouTube, is posted. Luo et al. developed the negative WOM , which impacts on negative firms [15,16], and chintagunata et al. developed the impact of customer purchases based on online reviews [17]. Table 1 shows further research methods in the field of data analysis and machine learning. Big data describe the large and complex collection of datasets that are difficult to process. Typically, big data explain the dataset based on the extracted features from the information [18]. The features are related to data scale and quantity. Similarly, the generated velocity shows the rate and speed at which data is analyzed. On the other hand, the data variety shows the structured or unstructured data formats in an additional characteristic like value, veracity, variability, and visualization [19]. Big data alone is not the exact answer for data consideration but in the case of internet businesses, it can consider raw data for more transformation [20]. Generally, extracting the hidden part of the collected information to create the business knowledge needs to understand the B2B environment [21]. User, big data analysis refers to store, acquire, process and analyze the user's velocity, volume, variety related data with the goal of making the meaningful information for decision making on firms and discover the novel sights for the business area in social media websites [22]. User relationship management is one of the important pieces of information in most organizations. In B2B marketing, managing the user's relationship has the fact to extract useful information and increase the insights of the dataset. E.g., the combination of web servers and big data will increase the user's

Table 1. Big data and machine learning related researches

Research	Used Dataset	Proposed Method	Contribution
Zhang et al. (2011) [24]	- Users messages (868000) - Online reviews (8226)	Text mining	Extracting the product information and applying the market structure analysis
Wu et al. (2014) [25]	- Users product reviews (350000) - Shared posts (115000)	- Latent Dirichlet Allocation - Sentiment Analysis	Extracting the users satisfaction
Netzer et al. (2015) [26]	- Communication	Conceptual	Content analysis and problem discussion about overcoming the issue
Morstatter et al. (2016) [27]	- Restaurant reviews (696) -200 brands in twitter	- Latent Dirichlet Allocation - Text mining	Sentence and word-based comparison
Meng et al. [28]	- 3years users dataset (250000)	Time-varying model	Test the effectiveness of the time-varying model
Li et al. [29]	- Healthcare	Content analysis	Big fata analysis in linguistic theory
Wang et al. [30]	Healthcare	Content analysis	Big data analysis capability
Minnick et al. [31]	- Twitter (18 million) - Online reviews (5830)	Text mining	Users generated content User feedback prediction

relationship management to generate better product recommendations and make the competition in the line of product generating system [23].

3. Proposed System and Results

This section presents the proposed big data processing. The main contribution of this section is divided into five parts. Data collection, data pre-processing, data visualization, data aggregation, and data analysis. We discuss all the mentioned parts in detail below. Figure 2 shows the conceptual model of the proposed system architecture. The presented conceptual diagram contains information related to users in social media content. To improve the sale growth, finding the relationship between users, finding the users' preferences, and similarly controlling the variables based on the collected dataset related to social media users is the reason to increase the sales in various websites.

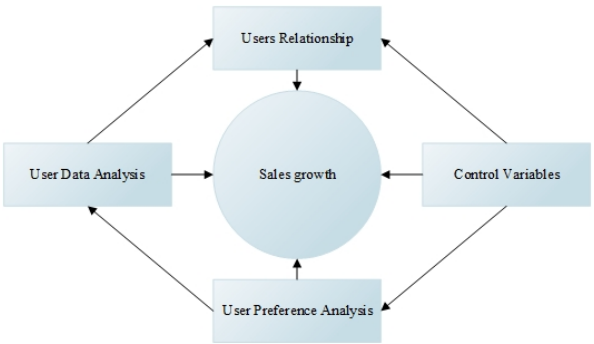

Figure 2. Conceptual Diagram

3.1. Data Collection

The first step to start the analysis of a project is to collect the dataset from the focused topic. In this paper, the dataset is related to various categories in the Twitter website related to users' interest in topics and sharing information on social media. The number of likes that show the users' interest in that topic. The history of sharing the information and mostly retweeting the information causes the topic to be available for more users. Table 2 represents the summary of the collect dataset. As is

Table 2. Summery of twitter data collection

Category	Tweet	Retweet	Reply	Likes
Study	5.331.906	1.317.568	4.174.749	544.346
Fashion	1.906.538	272.199	647.262	80.971
Lifestyle	51.445	14.377	16.771	884
Research	574.555	128.484	264.919.577	193.641
Job	113.283	42.525	1.909.153	11.577
News	1.468.579	334.541	560.213	1.090.153
Shopping	39.838	364.895	17.937	560.213

shown in the table, the dataset is divided into seven categories. The study, fashion, lifestyle, research, job, news, and shopping. The rest is divided into four sections which show the history of the tweeting, retweeting, replying and liking.

3.2. Data Pre-Processing

Data pre-processing step is an essential part of the data analysis process, but sometimes it can be under-emphasized. The first step to start the pre-processing is to remove the bot-created tweets from the dataset because this information is not from a real user and can affect the final result of the process. Similarly, the spam messages and accounts were deleted too. To get to know about the spam accounts , by checking the duplicate messages in the account, we can recognize that it's spam. In our process, by using the programming abilities, the spam accounts and bot messages are classified based on several features, including the writing style, account followers, the domain of topic, and the shared knowledge. In summary, data pre-processing in this system involves removing spam tweets, bot accounts, available hashtags, unnecessary URLs, grammatical mistakes, and stop words. The processed tweets are defined as 73 million tweets to increase the data quality and the analysis results, which is based on users' preferences. After the presented process, 50 million brands related to 407 companies extracted.

3.3. Data Visualization

Data visualization is a popular step to get the main idea of the dataset. It is the main technique, to sum up, the big data graphics. To do this, word cloud and sentiment analysis were applied to the proposed system. The word cloud technique identifies the frequency of the word in the dataset. To do this, it follows the nouns, adjectives, adverbs , and verbs in a document. To normalize the distribution matrix of the words in a document, the R software generates the word cloud. Figure 3 presents the word cloud overview in this system. It is indicating the highest frequency of the major service to users.

Figure 3. Overview of the Word Cloud

Sentiment analysis in the proposed system contains the Latent Dirichlet Allocation procedure to extract the user's interest. To explore more detailed information, a supervised learning approach was applied to visualize and summarize the dataset. Figure 4 presents the visualization process based on

the LDA system. The extracted topic and the highest related words are all presented. The result of the Figure 4 is based on the relationship between 46 various Twitter accounts; the number of retweets and likes has the highest effect on data cationization. Overall, it is shown the great information spread between users.

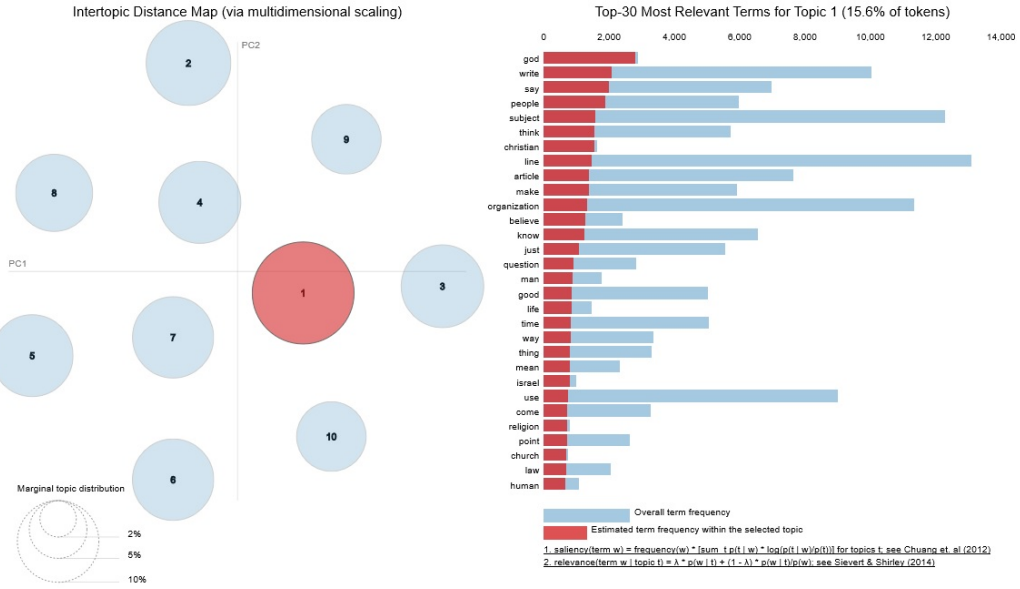

Figure 4. Data Visualization Process Using Big Data and Machine Learning

3.4. Data Aggregation

Based on the data collected from various sources, it is a must to integrate the dataset. In this process, we applied the natural language processing (NLP) technique on 61 million tweets and stored this data in an SQL database with the categories information, the merged tweets, and the finalized dataset. The database relationships are based on digital marketing and marketing analysis in big data. The big data key transforming technique to manage the formats is using the various temporal intervals. In the proposed system, the aggregation process is defined as multiple yearly and months which contain the 407 brands. Figure 5 presents the data aggregation process in this system. Based on extracting relevant keywords from the tweeter API, the related contents and tweets collected and fetch into the aggregation process. The aggregation process extracts the main opinion of the shared information between users.

3.5. Data Analysis

To finalize the data analysis process, 407 companies and their activities on Twitter is evaluated based on a monthly basis. Each company's starting date of sharing information and having a Twitter account is different. This shows the unbalanced panel information. To properly analyzing their information, we need to extract the unique advantages of each company. This process causes casual analysis based on data-controlled variables. Since the extracted information for each company is disparate in multiple perspectives, e.g., organization culture, leadership style, the size of the firm, etc. The ability to use each company's special control information is valuable. Table 3 presents the discriminant validity of the analyzed dataset. The divided categories show the average of user's activity analysis on Twitter based on various companies' brands. The categories are divided into user's big data analysis, user's preference relationship, and the analytic culture.

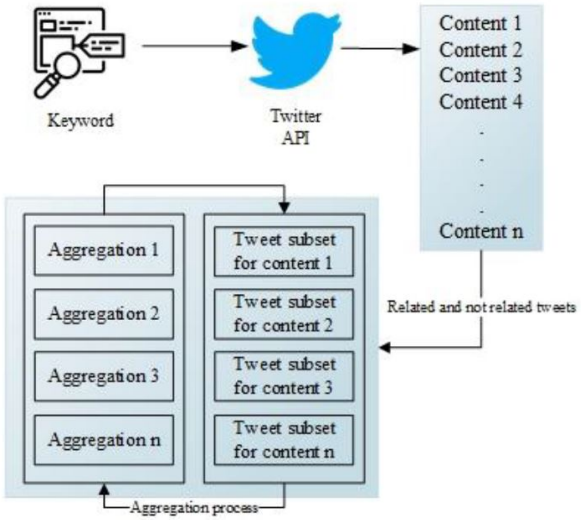

Figure 5. Data Aggregation Process in Twitter API

Table 3. Summery of twitter data collection

#	Aerage	1	2	3
Users big data analysis	0.618	0.743	-	-
Users preference relationship	0.413	0.078	0.616	-
Analytic culture	0.645	0.507	0.012	0.769

3.6. Development Environment

This section presents the development environment in detail. Experimental setup shows in Table 4. All experiments and results of the system are carried out using Intel(R) Core (TM) i7-8700 CPU @3.20GHz processor with 32 GB memory. The library and framework used in the proposed system are Jupyter notebooks. The programming language used in the designing of this system is WinPython–3.6.2.

3.7. Topic Clusters Segregation

Topic cluster segregation is multiple fragments of the contents which share the related topics and subtopics. This process covers the special subjects. Similarly, it makes it easy to understand the total process of contents. This process personalizes the search result based on the extracted keywords and makes the search engine understand the semantic analysis in related concepts. Similarly, this process provides trust to users and shows trustworthy results. Figure 6 presents segregated topics in the proposed system. The importance of topic cluster collection is to categorize the similar subtopics cluster into one main category.

Table 4. Development environment of the proposed system

Component	Description
Programming Language	SQL Server
Operating System	Windows 10, 64bit
Browser	Google Chrome, Opera
CPU	Intel(R) Core (TM) i7-8700 CPU @3.20GHz
Memory	32GB

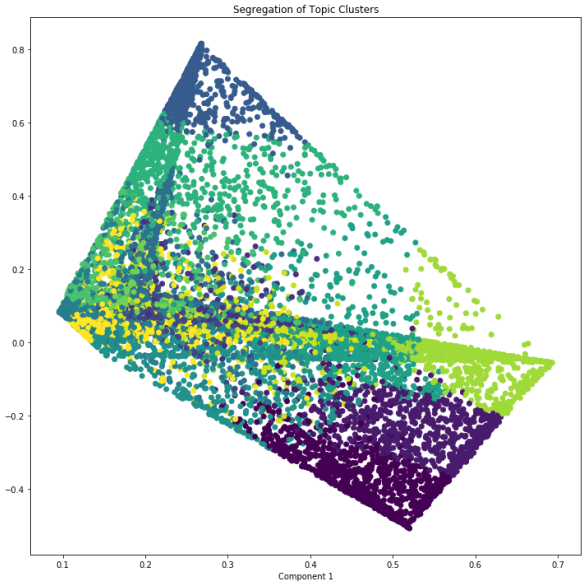

Figure 6. Topic Cluster Segregation

4. Conclusions

Big data analysis on user-generated content contains unnecessary information related to users' activities on social media platforms. Moreover, big data can make the firms' operation possible and, similarly, activate the business. The major challenges in this process cause the researchers to avoid taking advantage of dataset. The proposed model is presenting the importance of B2B system on a stock dataset. The B2B literature illustrating the related social media firms, the user's online activities, and the extraction of valuable information to understand the user's preferences for prediction. Most of the B2B user's dataset on firms has not enough information related to users' activities. To overcome this issue, user-generated content has significant access to B2B firm's performance.

Acknowledgments: This work was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0016977, The Establishment Project of Industry-University Fusion District)

References

1. How, S.M.; Lee, C.G. Customer satisfaction and financial performance-linear or non-linear relationship: a case study of Marriot International. *Current Issues in Tourism* **2021**, *24*, 1184–1189.
2. Allison, P.D. *Fixed effects regression models*; SAGE publications, 2009.
3. Armbrust, M.; Xin, R.S.; Lian, C.; Huai, Y.; Liu, D.; Bradley, J.K.; Meng, X.; Kaftan, T.; Franklin, M.J.; Ghodsi, A.; others. Spark sql: Relational data processing in spark. Proceedings of the 2015 ACM SIGMOD international conference on management of data, 2015, pp. 1383–1394.
4. Mahendra, M.; Luo, Y.; Mills, H.; Schenk, G.; Butte, A.J.; Dudley, R.A. Impact of Different Approaches to Preparing Notes for Analysis With Natural Language Processing on the Performance of Prediction Models in Intensive Care. *Critical care explorations* **2021**, *3*.
5. Bjeladinovic, S. A fresh approach for hybrid SQL/NoSQL database design based on data structuredness. *Enterprise Information Systems* **2018**, *12*, 1202–1220.
6. Rabetino, R.; Kohtamäki, M.; Brax, S.A.; Sihvonen, J. The tribes in the field of servitization: Discovering latent streams across 30 years of research. *Industrial Marketing Management* **2021**, *95*, 70–84.
7. Zhang, Z.; Zhang, D. What is Data Science? An Operational Definition based on Text Mining of Data Science Curricula. *Journal of Behavioral Data Science* **2021**, *1*, 1–16.

8. Shahbazi, Z.; Byun, Y.C. LDA Topic Generalization on Museum Collections. In *Smart Technologies in Data Science and Communication*; Springer, 2020; pp. 91–98.
9. Great, S. Research Methods. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing* **2021**, p. 194.
10. Kushwaha, A.K.; Kumar, P.; Kar, A.K. What impacts customer experience for B2B enterprises on using AI-enabled chatbots? Insights from Big data analytics. *Industrial Marketing Management* **2021**, *98*, 207–221.
11. Dhanji, N.; Brouwer, W.; Donaldson, C.; Wittenberg, E.; Al-Janabi, H. Estimating an exchange-rate between care-related and health-related quality of life outcomes for economic evaluation: An application of the wellbeing valuation method. *Health Economics* **2021**.
12. Erevelles, S.; Fukawa, N.; Swayne, L. Big Data consumer analytics and the transformation of marketing. *Journal of business research* **2016**, *69*, 897–904.
13. Fan, S.; Lau, R.Y.; Zhao, J.L. Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Research* **2015**, *2*, 28–32.
14. Farrell, A.M. Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of business research* **2010**, *63*, 324–327.
15. Frizzo-Barker, J.; Chow-White, P.A.; Mozafari, M.; Ha, D. An empirical study of the rise of big data in business scholarship. *International Journal of Information Management* **2016**, *36*, 403–413.
16. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management* **2015**, *35*, 137–144.
17. Chintagunta, P.; Hanssens, D.M.; Hauser, J.R. Marketing science and big data, 2016.
18. Chintagunta, P.K.; Gopinath, S.; Venkataraman, S. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing science* **2010**, *29*, 944–957.
19. Choi, I. Unit root tests for panel data. *Journal of international money and Finance* **2001**, *20*, 249–272.
20. Chu, Z.; Gianvecchio, S.; Wang, H.; Jajodia, S. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on dependable and secure computing* **2012**, *9*, 811–824.
21. Culotta, A.; Cutler, J. Mining brand perceptions from twitter social networks. *Marketing science* **2016**, *35*, 343–362.
22. Van Der Meer, T.G. Automated content analysis and crisis communication research. *Public Relations Review* **2016**, *42*, 952–961.
23. Devi, D.N.; Kumar, C.K.; Prasad, S. A feature based approach for sentiment analysis by using support vector machine. 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE, 2016, pp. 3–8.
24. Zhang, Z.; Zou, Y.; Gan, C. Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. *Neurocomputing* **2018**, *275*, 1407–1415.
25. Wu, P.J.; Lin, K.C. Unstructured big data analytics for retrieving e-commerce logistics knowledge. *Telematics and Informatics* **2018**, *35*, 237–244.
26. Netzer, O.; Feldman, R.; Goldenberg, J.; Fresko, M. Mine your own business: Market-structure surveillance through text mining. *Marketing Science* **2012**, *31*, 521–543.
27. Morstatter, F.; Wu, L.; Nazer, T.H.; Carley, K.M.; Liu, H. A new approach to bot detection: striking the balance between precision and recall. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016, pp. 533–540.
28. Meng, X.; Bradley, J.; Yavuz, B.; Sparks, E.; Venkataraman, S.; Liu, D.; Freeman, J.; Tsai, D.; Amde, M.; Owen, S.; others. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research* **2016**, *17*, 1235–1241.
29. Li, X.; Yuan, J.; Ma, H.; Yao, W. Fast and parallel trust computing scheme based on big data analysis for collaboration cloud service. *IEEE Transactions on Information Forensics and Security* **2018**, *13*, 1917–1931.
30. Wang, A.H. Detecting spam bots in online social networking sites: a machine learning approach. IFIP Annual Conference on Data and Applications Security and Privacy. Springer, 2010, pp. 335–342.

31. Minnick, K.; Noga, T. The influence of firm and industry political spending on tax management among S&P 500 firms. *Journal of Corporate Finance* **2017**, *44*, 233–254.