An investigation of the relationship between Privacy Crisis, Public Discourse on Privacy, and Key Performance Indicators at Facebook (2004–2021)

ABSTRACT

We use Facebook as a case study to investigate the complex relationship between the firm's public discourse (and actions) surrounding data privacy and the performance of a business model based on monetizing user's data. We do so by looking at the evolution of public discourse over time (2004–2021) and relate topics to revenue and stock market evolution Drawing from archival sources like Zuckerberg We use LDA topic modelling algorithm to reveal 19 topics regrouped in 6 major themes. We first show how, by using persuasive and convincing language that promises better protection of consumer data usage, but also emphasizes greater user control over their own data, the privacy issue is being reframed as one of greater user control and responsibility. Second, we aim to understand and put a value on the extent to which privacy disclosures have a potential impact on the financial performance of social media firms. There we found significant relationship between the topics pertaining to privacy and social media/technology, sentiment score and stock market prices. Revenue is found to be impacted by topics pertaining to politics and new product and service innovations while number of active users is not impacted by the topics unless moderated by external control variables like Return on Assets and Brand Equity.

Keywords: public discourses; data protection; social media; privacy; topic modelling; business model; financial performance

INTRODUCTION

Data privacy is a controversial topic because it is an ambiguous construct. Data is managed and usually owned by digital platforms while privacy relates to the intimacy of users. Social media platforms in particular thrive on users' data. No other firm better than Facebook illustrates this tension between monetization and respect of users' data as well as the recurrent privacy crisis inherent to such business model (e.g. 2008, 2012, 2015, 2018, 2020). With 3 billion users, it is one of the world's largest aggregators of user data. According to Forbes, the Facebook brand is valued at $70.39 billion (fifth most valuable in the world) in 2020, generating $117,929 billion in revenue and $ 39,370 billion in net income in 2021 [1]i.

Yet, research related to data privacy has focused predominantly on user behavior and the related privacy paradox (Gerber et al., 2018; Barth and de Jong, 2017; Kokolakis, 2017, Martin et al, 2017; Smit et al., 2014; Afroz et al., 2013; Lee et al., 2013; Debatin et al., 2009). Few studies have looked at online privacy from the business perspective, with some notable exceptions. Stutzman et al.,(2011) and Pollach(2005) have explored the use of language by corporations in their official privacy policies and found that these formal policies can "obfuscate, enhance and mitigate unethical data handling practices" (Pollach, 2005, pp. 221). However, no studies have investigated the relationship between privacy discourse, public opinion crisis and the success of digital business as measured by Key Performance Indicators such as numbers of active users, revenue and stock market prices. This will help us to answer the following research questions:

[1] Source: https://www.statista.com/statistics/277229/facebooks-annual-revenue-and-net-income/ as accessed on 10/02/2022

- What is the corporate public discourse around privacy?
- What is the relationship between public discourse during privacy crisis and Key Performance Indications (active users, revenue and stock market prices?)

Data Collection and Research Methodology

The main motive of this research is to understand the thematic evolution of Facebook transcripts with time and see how those themes (in particular the ones related to privacy) relate to key performance indicators of the business model..

The detailed research methodology is illustrated below:

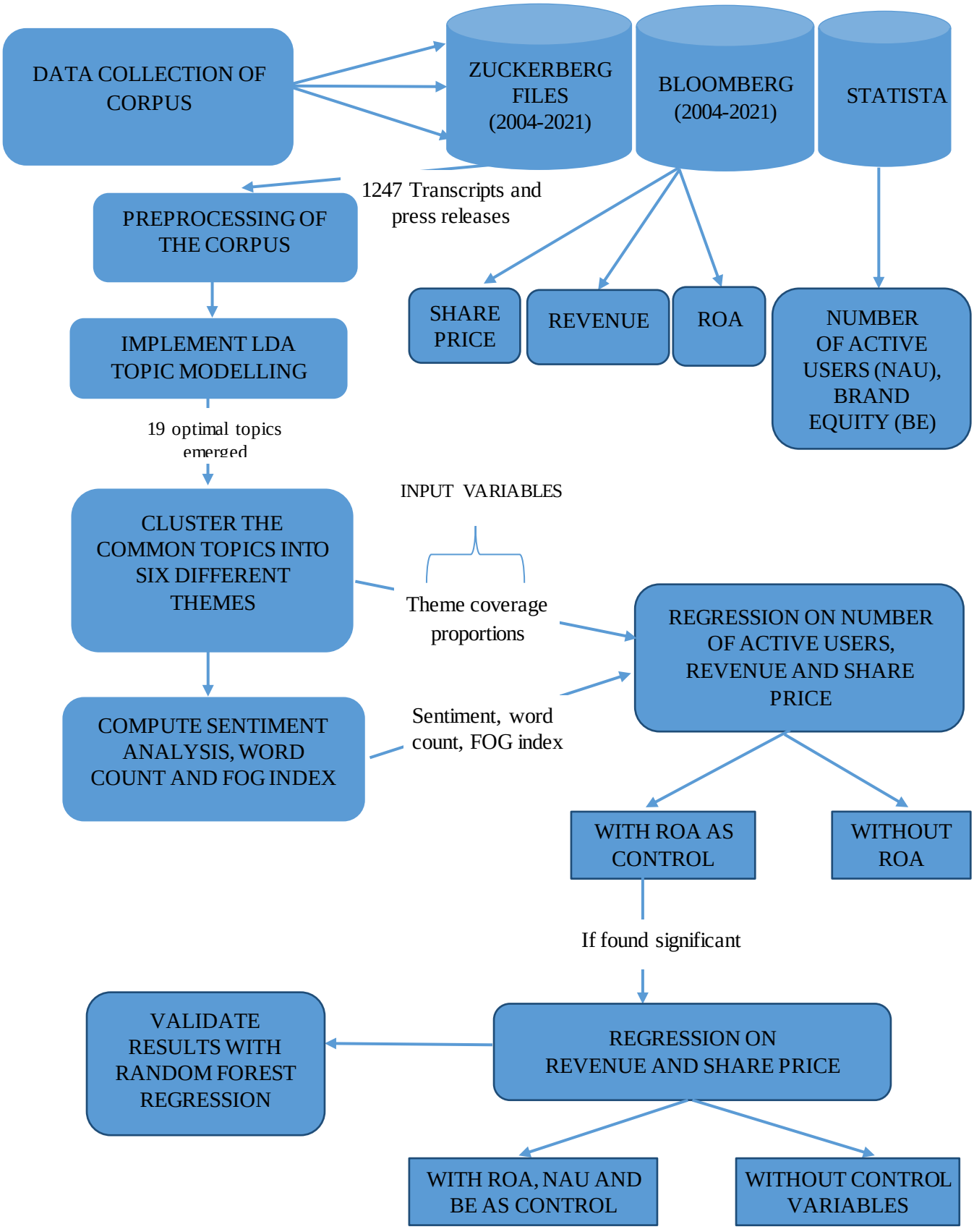


Figure 1. Research Methodology

The Zuckerberg[2] files between the years 2004 to 2021 are collected and filtered in terms of "Transcripts" to extract only the files which are transcripts of Mark Zuckerberg's speeches. The search retrieved 1247 transcripts over the years.

The LDA topic modelling algorithm (Thielmann et al.,2021) was implemented to identify the major topics and themes covered in the transcripts. This would help understand which themes have evolved in importance for Facebook.

Sentiment analysis was performed in R (Wang et al., 2022) to extract the sentiment of the Zuckerberg transcripts in order to understand how the announcements are strategically framed around issues of privacy and the impact on changing business models.

The word count of each article is also computed and incorporated as a predictor in the dataset to understand how the degree of conciseness of a transcript/article would impact the performance of the firm.

An additional metric to the degree of conciseness is also the rate of complexity and readability of the Zuckerberg announcements. This would provide an insight into how firms strategically use confounding and high complexity words without providing a clear assurance to the users about their stand on privacy. For this purpose, the metric FOG index (Ahmadi et al.,2021) is computed in R. The transcripts have information that encompasses various Facebook missions, key company events, the launch of new social media technologies, stand on governance, the launch of new products and services, and the stand on tackling user privacy issues for the growth of the business. With such vast topics under consideration, the textual data of these announcements consist of numbers, weblinks and special characters, and several stop words. The transcripts downloaded in "pdf" format are initially read into RStudio using the predefined package "pdftools". The transcripts are stored in a "Corpus" object to enable further textual processing. In order to identify the major topics in these announcements, topic modeling is applied which is an unsupervised algorithm that clusters the announcements based on the similarity between them. As topic modeling works on understanding the word frequency and the distance between words it's important to exclude the stopwords, punctuation symbols, numbers, and other special characters. The transformed corpus is converted into a Document term matrix. This is a format that stores the words and their corresponding frequency of certain terms concerning the documents. It is crucial to identify the optimal number of topics to segregate the topics and themes accordingly. A coherence score plot is constructed to find the optimal number of topics between 1 to 20 and the highest coherence score was found to occur at 19. The next use case of the algorithm was to map each document to its predominant topic cluster. Further, the top 15 frequently occurring terms in each topic were extracted to observe the context in each topic. Further analysis has revealed that there are 6 emergent themes among the 19 topics.

[2] https://epublications.marquette.edu/zuckerberg_files/

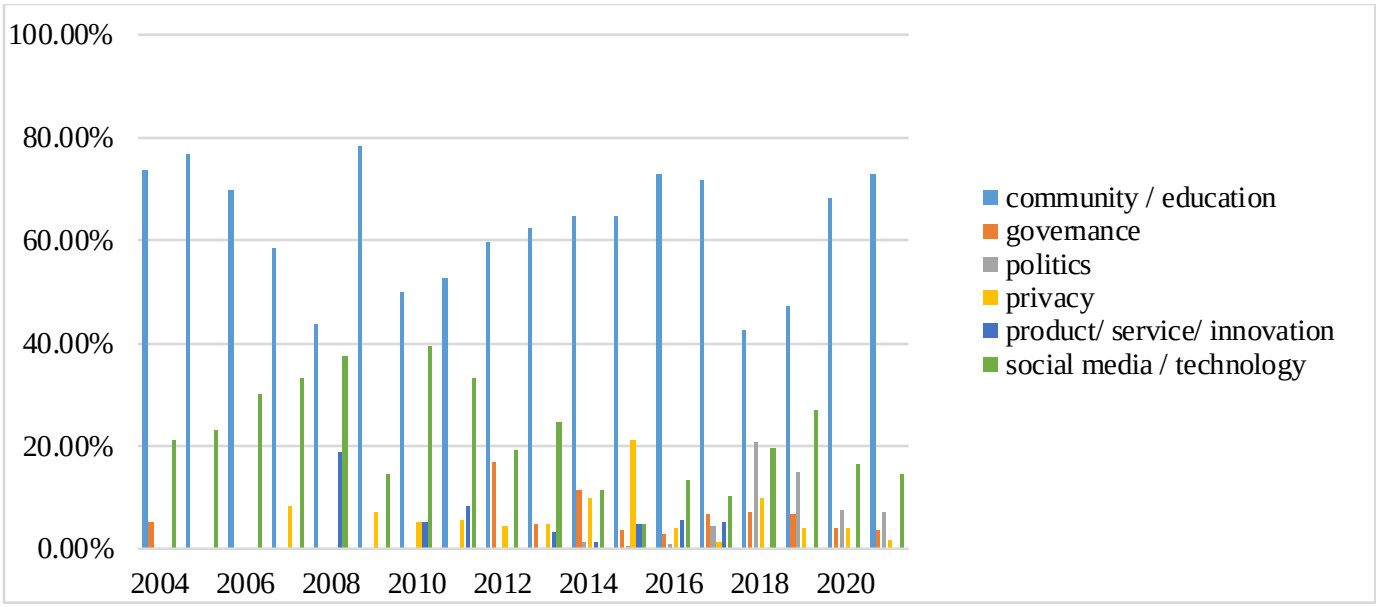
Results and Discussion

Table 1 below describes the themes and the topics under each theme. These themes when mapped to the individual transcripts to codify the apt domain under discussion in each of them.

Table 1. Themes and topics classification and description

Theme	Mapped Topics	Description
Community and social communication	2,12,14,15,18	Focuses mostly on the vanity statement and contribution of Facebook to funding schools, charities, and social initiatives
governance	1,5	Transcripts on board of governance decisions, shareholder meetings, earnings call transcripts, and conference calls
Politics	6	Political viewpoints, electorate information, and freedom of expression
Privacy	8,9,13	User privacy concerns, accountability, and need for transparency
product/ service/ innovation	7,10	Launch of new products, services, and technologies
social media technology	3,4,11,16,17,19	Digital breakthroughs, social media communities, and the launch of augmented Virtual Reality

The results of the thematic evolution and topic-wise evolution over the years 2004-2021 are illustrated below:



. Figure 2. Thematic evolution (2004-2021)

Figure 2 illustrates the evolution of the six themes over the period. The themes of "Community/Education" and "Social media/technology" are found most dominant while other themes like politics, governance, and privacy are found to be emerging in terms of frequency of discussion in the transcripts. This implies that Facebook needs to reiterate a stronger stand on important issues like securing user privacy and mitigating the spread of adverse political opinions and posts.

Similarly, another thematic evolution analysis without the confounding influence of the dominant theme "Community/Education" and with respect to the three themes "government", "politics" and "privacy" is illustrated in Figure 3:

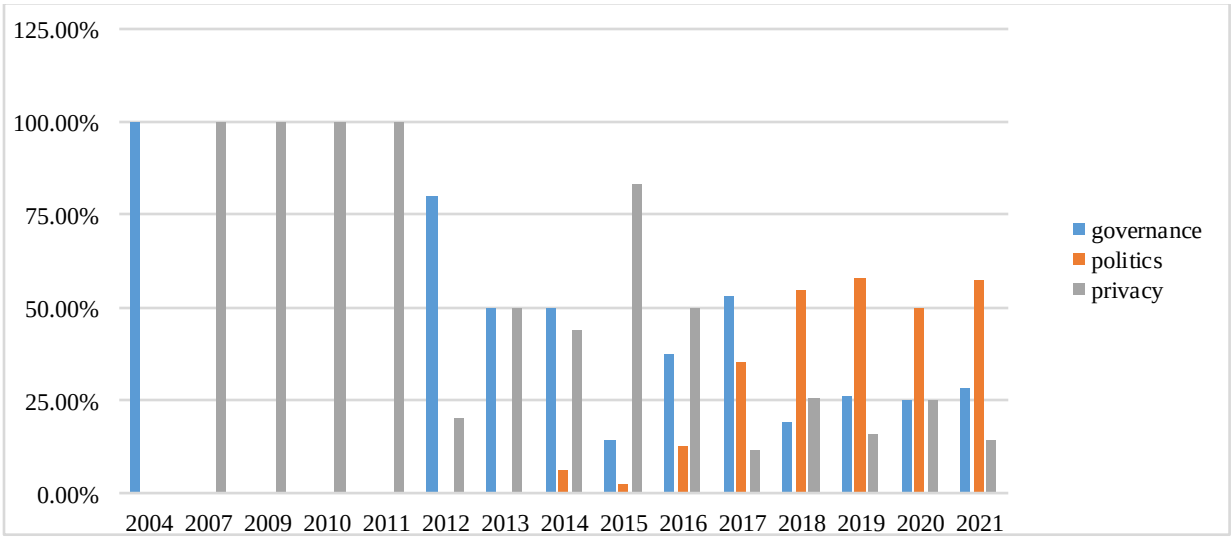


Figure 3. Thematic evolution for the themes "government", "politics" and "privacy" (2004-2021)

The themes of "governance" and "privacy" are found most dominant while other themes like politics are found to be emerging with time in terms of frequency of discussion in the

transcripts. Similarly, the evolution of the different 19 topics with time are illustrated in Figure 4 while the topics pertaining to only governance, politics and privacy themes are illustrated in Figure 5:

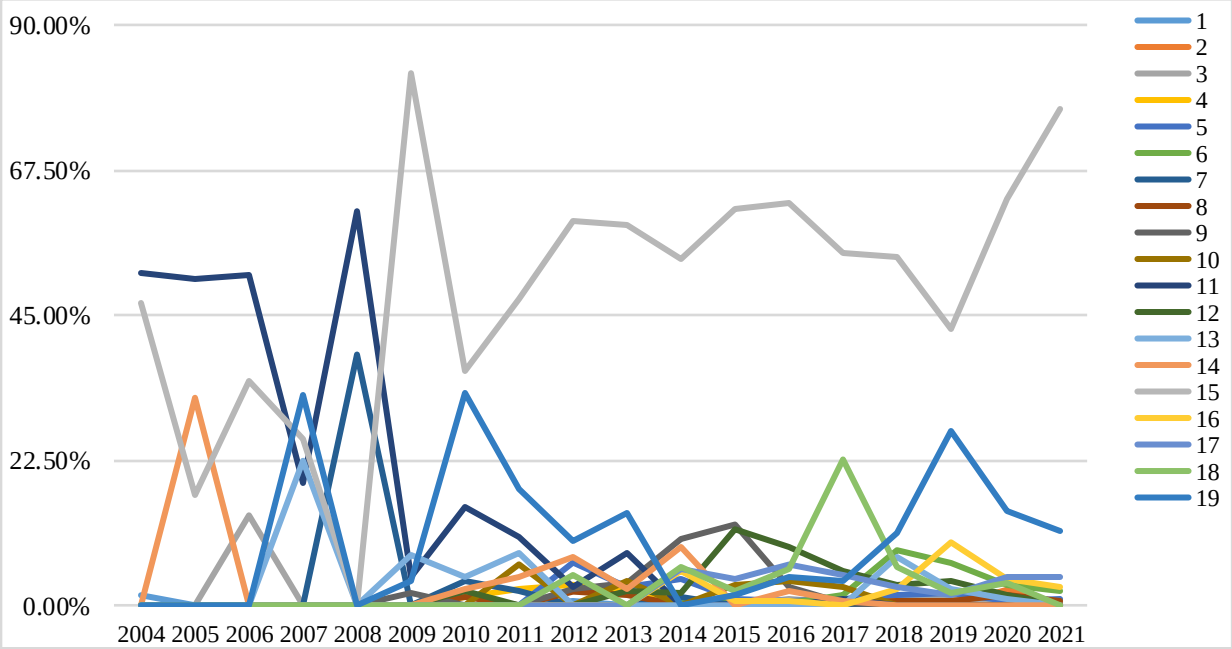


Figure 4. Topic cluster evolution (2004-2021)

As illustrated in Figure 4, topic 11 (on social media interviews conducted with Mark Zuckerberg), topic 19(social media opinions), and topic 15 (related to philanthropic and charitable initiatives) are the most dominant of the themes. These corroborate the thematic analysis above suggesting that topics classified under "Community/Education" (topic 15) and "Social Media" (topics 11 and 19) are the most frequently discussed strands in Zuckerberg's speech. There is scope for privacy-related topics like "strategy/innovation", "litigation" and "human rights/ global development" to be discussed in greater detail to provide better assurance to users about their private information being secured and ensuring everyone receives fair access to the Internet.

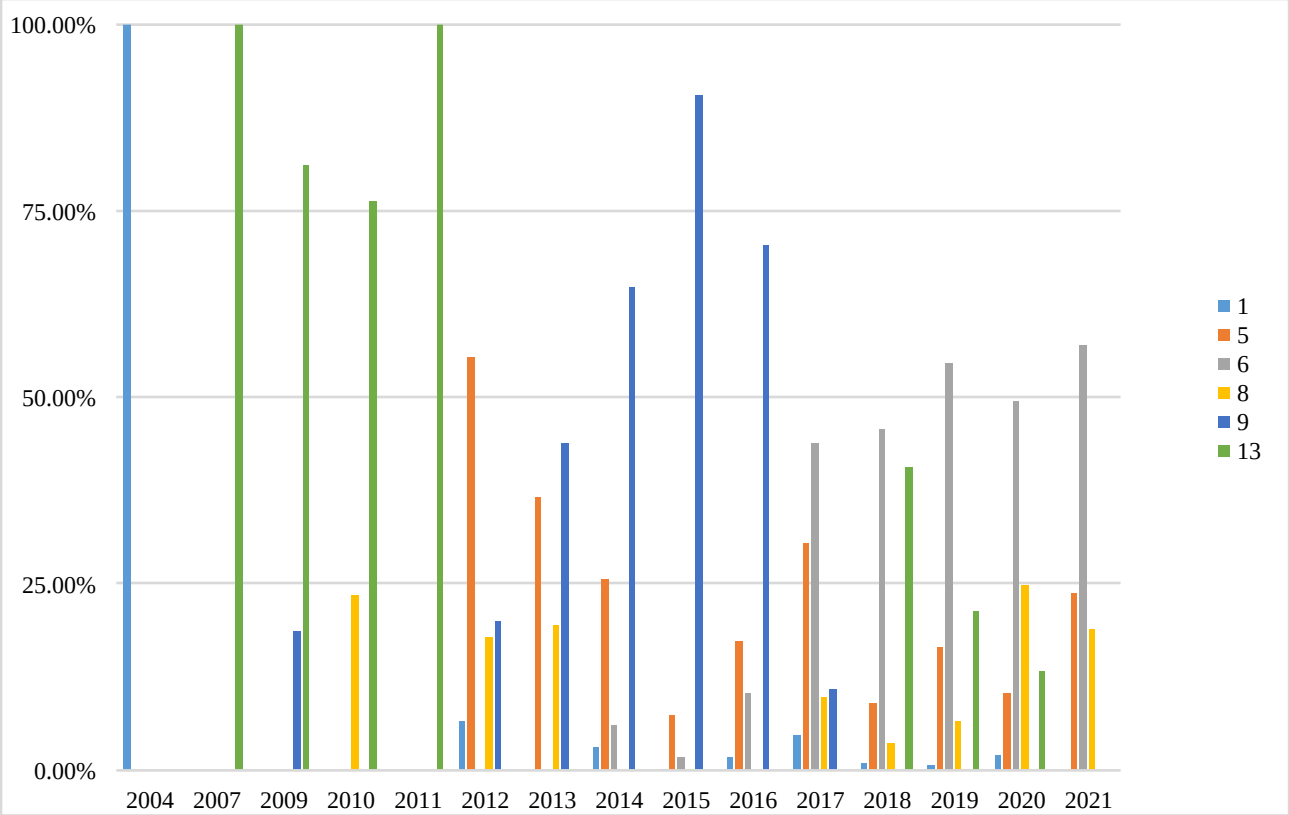


Figure 5. Topic cluster evolution for topics under governance, politics and privacy (2004-2021)

Reconstructing the above Figure 4 after limiting the topics to only 1,5,6,8,9 and 13 in Figure 5, it is still found that topic 9 (privacy related) are the most dominant followed by topic 6 (politics) and topic 5(governance). This implies that Facebook is now attempting to strengthen its ground on securing user privacy controls.

To identify the favourability of the subjected information towards the context sentiment analysis is performed in R using the 'qdap' package. Expressions leading to the positive side of the context indicate favourability while expressions with a negative connotation indicate the unfavorable nature of the content towards the context. In order to understand if the companies are strategically framing the announcements with more positive terms assuring less threat to user privacy, sentiment analysis is performed. The sentiment scores of the announcements obtained are chronologically provided to examine the disclosure patterns of companies with time.

Further, in order to analyze the complexity and readability of the documents in R, fog index is computed in R by the predefined package "korPus". The data files are initially tokenized into a standardized form and hyphenated for converting them into readable files. The files are then converted to readable files using a wrapper function which computes different complexity scores for the document. Of all the scores, the "grade" score is extracted and assigned as the FOG index score. This reflects how the complexity of the announcements made by social media companies are synchronized to the stand on user privacy and evolving business models. Wherever, the complexity is found to be high, this indicates a higher likelihood of manipulation of the disclosures to align to user needs. The financial indicators of the firm performance

measured by Share price, Revenue and Return on Assets are sourced from the Bloomberg database while the data on number of active users and brand equity is retrieved from the Statista database.

The thematic topic coverage proportions, sentiment, word count and FOG index are regressed on the outcome variables Share price, Revenue and the number of active users with and without the impact of control variable "Return on Assets". Similarly, if the variables are found to be significant, brand equity is introduced as an additional control variable. This is to introduce the moderating influence of number of active users and brand equity on the outcome variables Share price and Revenue. The regression results are validated by the random forest regression machine learning algorithm. The results are illustrated below:

The regression results on Share price, Revenue and Number of Active users as outcome variables without any control variable are illustrated below in Table 2:

Table 2. Summary of regression model results on Share price, Revenue and Number of Active users as outcome

	Model 1	**Model 2**	**Model 3**
Variables	**Coefficients**	**Coefficients**	**Coefficients**
governance	0.21(0.636)	-882(0.236)	-0.034(0.858)
community/education	0.51(0.12)	-193.24(0.721)	-0.005(0.97)
politics	0.1 (0.83)	-1459.4(0.048)	0.522(0.011)
privacy	0.48 ***(0.026)	-82.6(0.907)	-0.027(0.88)
product/service/innovation	0.08(0.882)	538.4 ***(0.003)	-0.002(0.992)
social media/technology	0.074***(0.42)	-545.3(0.358)	0.097(0.523)
Word count	0.17 (0.743)	3354 ***(0.0004)	0.185(0.412)
FOG index	0 (0.89)	-47.5***(0.004)	0(0.927)
Sentiment_score	0.21*** (0.009)	-133.02(0.503)	0.07(0.191)

The three models Model 1, Model 2 and Model 3 imply the regression results for the key performance indicators Share price, Revenue and Number of active users respectively. For the first regression model with the KPI Share price as dependent variable, the topics pertaining to privacy, social media/technology and sentiment score are found to be the most significant positive drivers for share price. This implies that providing assurance to users about privacy, latest technological trends and a more positive linguistic framing of transcripts has an impact on boosting the share price of Facebook. Similarly, for the second model with Revenue KPI as dependent variable, the news pertaining to politics, product/service/innovation news, word

count and FOG index are found significant. This indicates that keywords related to political agendas and new products/services, the number of words in transcript and degree of complexity of transcript are important drivers of revenue. While, product and service-related news and word count are positive drivers, the news pertaining to politics and FOG index are negative drivers. This is because topics pertaining to politics may be controversial and detrimental to the reputation of Facebook in the wake of the Cambridge Analytica scandal and the use of complex words may cloud decision-making and cause a dip in revenue. However, political news is found to have a positive impact on the number of active users KPI as illustrated in Model 3. Overall, the topic-related variables privacy, social media/technology and sentiment score are found to be significant implying that the transcripts pertaining to privacy and social media and the overall sentiment have an impact on the key performance indicators of the firm.

Further, to improve the explainability of the model, other financial control variables like Profitability, Return on Assets and Price Earnings ratio have also been incorporated. However, Profitability and Price Earnings ratio were found to be highly correlated (>80%) and eliminated therefore, ROA is the only control variable considered.

With control variable ROA:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-5.374e+01	5.740e+01	-0.936	0.3494	
governance	1.033e+02	6.064e+01	1.703	0.0889	.
community.education	1.481e+02	5.789e+01	2.558	0.0107	*
politics	2.938e+01	6.112e+01	0.481	0.6308	
privacy	3.780e+01	6.151e+01	0.615	0.5389	
product.service.innovation	-1.812e+01	6.561e+01	-0.276	0.7825	
social.media.technology	4.944e+01	5.634e+01	0.878	0.3803	
Word.count	3.536e-04	3.523e-04	1.004	0.3158	
FOG.index	2.619e-02	5.795e-02	0.452	0.6513	
Sentiment_score	4.289e-01	5.593e+00	0.077	0.9389	
Return.on.Assets	8.098e+00	2.779e-01	29.147	<2e-16	***

Residual standard error: 53.24 on 1093 degrees of freedom
Multiple R-squared: 0.4886, Adjusted R-squared: 0.4839
F-statistic: 104.4 on 10 and 1093 DF, p-value: < 2.2e-16

Figure 6. Regression results for stock price as outcome

For stock price as outcome, news pertaining to community and education and ROA is found to be significant for the KPI stock price. This implies that with addition of control variable ROA, the topics related to community/education are the only significant themes driving the share price of Facebook with previously found significant topics related to privacy and politics confounded

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-9.843e+03	6.057e+03	-1.625	0.10445	
governance	6.738e+03	6.399e+03	1.053	0.29262	
community.education	1.580e+04	6.109e+03	2.586	0.00984	**
politics	2.014e+03	6.449e+03	0.312	0.75485	
privacy	5.369e+03	6.491e+03	0.827	0.40826	
product.service.innovation	-2.614e+03	6.924e+03	-0.378	0.70580	
social.media.technology	6.079e+03	5.945e+03	1.023	0.30670	
Word.count	2.802e-02	3.718e-02	0.754	0.45124	
FOG.index	2.734e+00	6.115e+00	0.447	0.65484	
Sentiment_score	1.520e+02	5.902e+02	0.258	0.79677	
Return.on.Assets	7.638e+02	2.932e+01	26.050	< 2e-16	***

Figure 7. Regression results for revenue as outcome

News pertaining to community and education and Return on Assets are significant for revenue as KPI. This implies that with addition of control variable ROA, the topics related to community/education are the only significant themes driving the revenue of Facebook with previously found significant topics related to politics and product/service/innovation confounded.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.157e+02	5.009e+02	-0.431	0.666852	
governance	1.320e+03	5.292e+02	2.494	0.012784	*
community.education	1.796e+03	5.051e+02	3.555	0.000394	***
politics	1.180e+03	5.333e+02	2.212	0.027145	*
privacy	1.295e+03	5.367e+02	2.412	0.016030	*
product.service.innovation	8.669e+02	5.726e+02	1.514	0.130283	
social.media.technology	1.185e+03	4.916e+02	2.411	0.016074	*
Word.count	-2.324e-03	3.074e-03	-0.756	0.449924	
FOG.index	-2.390e-01	5.056e-01	-0.473	0.636571	
Sentiment_score	-1.831e+01	4.881e+01	-0.375	0.707605	
Return.on.Assets	4.951e+01	2.425e+00	20.420	< 2e-16	***

Figure 8. Regression results for Number of active users as outcome

News pertaining to governance, community and education, politics, privacy, social media technology and ROA are found to be significant for number of active users as outcome which were not found significant earlier before incorporating ROA. Hence ROA is found to be a suitable significant moderator between topics highlighted and the number of active users.

This implies that suitably framing the news/transcripts with suitable privacy related assurance keywords and with the incorporation of ROA has an impact on boosting the financial performance of the firm.

Further, brand equity found to be significant is also incorporated as a control variable and the result is illustrated below for stock price and revenue as outcome variables respectively:

Stock price as outcome				
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.527e+01	4.773e+01	-0.948	0.34318
governance	2.346e+00	4.975e+01	0.047	0.96240
community.education	1.510e+01	4.861e+01	0.311	0.75613
politics	-2.931e+01	4.970e+01	-0.590	0.55545
privacy	-6.009e+01	5.068e+01	-1.186	0.23614
product.service.innovation	-7.295e+01	5.243e+01	-1.391	0.16452
social.media.technology	-4.026e+01	4.708e+01	-0.855	0.39272
Word.count	7.553e-04	2.633e-04	2.869	0.00423 **
FOG.index	1.012e-02	4.752e-02	0.213	0.83137
Sentiment_score	2.085e-02	4.586e+00	0.005	0.99637
Return.on.Assets	2.588e+00	3.440e-01	7.524	1.49e-13 ***
No.of.active.users.in.millions.	9.507e-02	3.523e-03	26.985	< 2e-16 ***
Brand.equity.in.millions...	-5.688e-05	1.279e-04	-0.445	0.65661

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 36.62 on 765 degrees of freedom				
Multiple R-squared: 0.7553, Adjusted R-Squared: 0.75				
F-statistic: 196.8 on 12 and 765 DF, p-value: < 2.2e-16				

Figure 9. Regression results for Stock price as outcome with ROA and Brand Equity as control variablesWord count, ROA and Number of active users are the only significant predictors of stock price with the effect of different topic-related news faded out as compared to the previous models with share price as KPI. Brand equity however, is not a significant control variable. This implies that brand reputation may not necessarily have an immediate impact on stock price unless moderated by a customer-centric variable.

Revenue as outcome				
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.733e+03	4.689e+03	-1.649	0.09949 .
governance	-5.061e+03	4.887e+03	-1.035	0.30077
community.education	1.032e+03	4.776e+03	0.216	0.82891
politics	-4.876e+03	4.882e+03	-0.999	0.31816
privacy	-6.116e+03	4.979e+03	-1.228	0.21968
product.service.innovation	-9.568e+03	5.150e+03	-1.858	0.06357 .
social.media.technology	-3.832e+03	4.625e+03	-0.829	0.40758
Word.count	5.951e-02	2.586e-02	2.301	0.02164 *
FOG.index	3.295e-01	4.668e+00	0.071	0.94375
Sentiment_score	1.336e+02	4.505e+02	0.297	0.76688
Return.on.Assets	2.008e+02	3.379e+01	5.942	4.27e-09 ***
No.of.active.users.in.millions.	1.022e+01	3.461e-01	29.544	< 2e-16 ***
Brand.equity.in.millions...	-3.851e-02	1.256e-02	-3.066	0.00225 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 3598 on 765 degrees of freedom				
Multiple R-squared: 0.7674,Adjusted R-squared: 0.7638				
F-statistic: 210.4 on 12 and 765 DF, p-value: < 2.2e-16				

Figure 10. Regression results for Revenue as outcome with ROA and Brand Equity as control variables

Word count, ROA, topic related to product and service related news, Number of active users and Brand Equity are significant predictors of number of active users KPI. This implies that brand reputation also boosts the revenue as it is moderated by an increase in number of active users.

References

Ahmadi, O., Louw, J., Leinonen, H., & Gan, P. Y. C. (2021). Glioblastoma: assessment of the readability and reliability of online information. *British Journal of Neurosurgery*, 1-4.

Thielmann, A., Weisser, C., Krenz, A., & Säfken, B. (2021). Unsupervised document classification integrating web scraping, one-class SVM and LDA topic modelling. *Journal of Applied Statistics*, 1-18.

Wang, W., Guo, L., & Wu, Y. J. (2022). The merits of a sentiment analysis of antecedent comments for the prediction of online fundraising outcomes. *Technological Forecasting and Social Change*, *174*, 121070.

[i] https://investor.fb.com/investor-news/press-release-details/2019/Facebook-Reports-Fourth-Quarter-and-Full-Year-2018-Results/default.aspx