Part A

This report aims to develop a strategy for traditional banks to compete with digital banks and improve customer satisfaction by analysing traditional and digital banks' customer reviews. It is a worldwide trend that digital-only banks get higher customer satisfaction than traditional banks, especially digital laggards (Lightico, 2022). The top 5 UK traditional banks (Santander, Barclays, HSBC, Nationwide, Natwest) and top 3 digital banks(Revolut, Monzo, Starling) by the number of users are selected as the subjects of review (Statista, 2022). Part A covers web scraping, data cleaning and Bag-of-Words model [Figure 1].

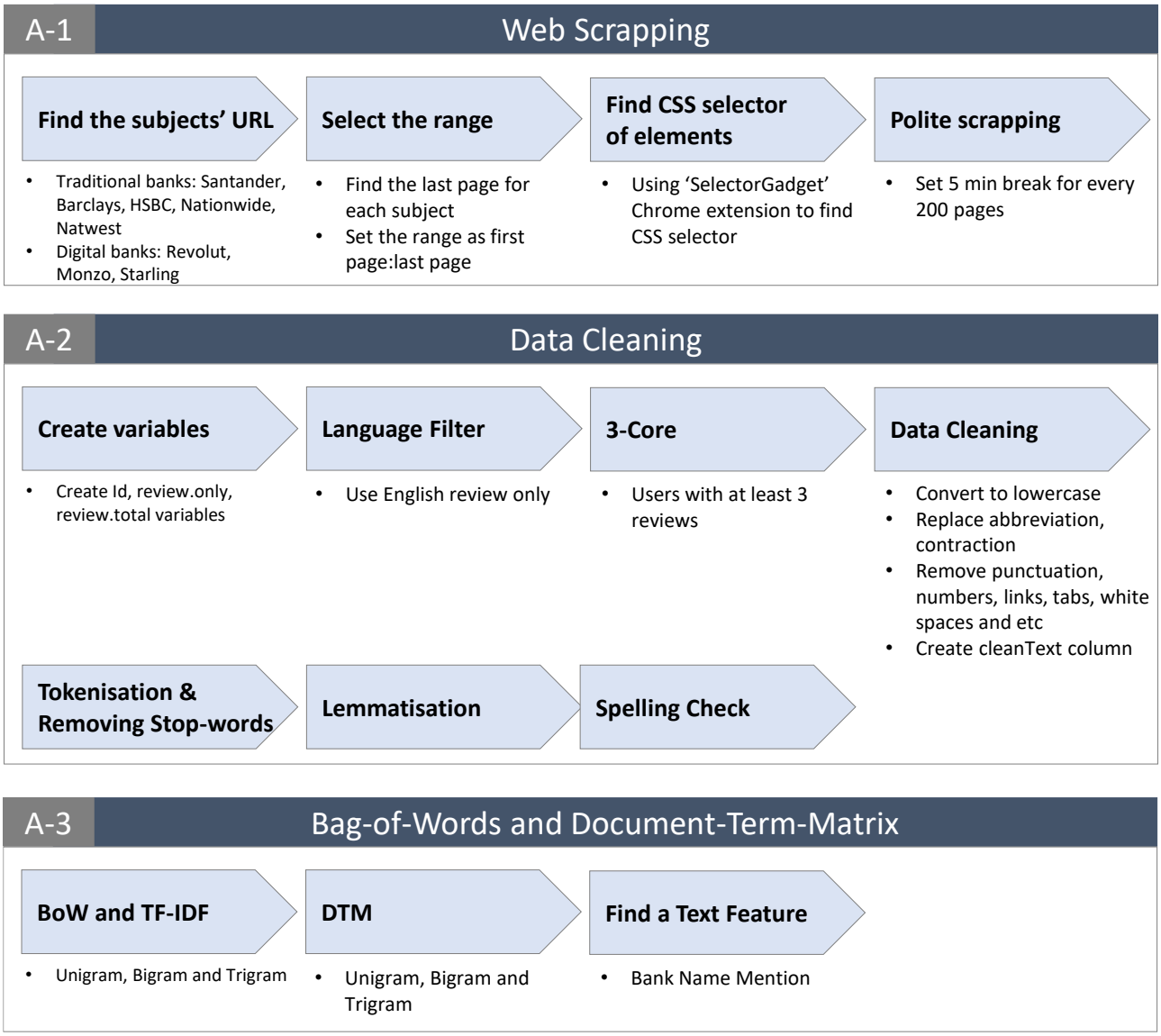


Figure 1 Part A Workflow

All reviews are scrapped from Trustpilot using rvest and xml2 packages (Trustpilot Reviews, 2022). The website is not scrapable using polite package; therefore, a five-minute break per every 200 pages is manually set for polite web scraping. A Chrome extension 'SelectorGadget' is used to generate CSS selectors for each element (Selectorgadget.com., 2022). One tricky part of scrapping is related to reviews only with title but without review content. If review content element is scrapped, the number of

review content and title are different. Therefore, title and upper element of title and review content are scrapped, then review content is extracted by removing the title from the upper element. From web scrapping, two datasets-traditional banks reviews and digital banks reviews- are generated and saved to one list named 'banks.data' for coding efficiency. Firgure 2 shows the examples of dataset and Figure 3 illustrates the composition ratio of dataset.

	company	cus_name	cus_tot_review	title	review	score
1	santander	jon bevan	28	santander are going broke-have held up...	santander are going broke- have held up my transfer of my own money which i was paying to a supplier since last thursday 31/03/22 and s till not released my funds !!!! i have over five thousand pounds with this so called bank and tried to pay online t hree thousand eight hundred pounds to a supplier and went online following all the rules but the so called frau d department stopped the transfer-spent six hours trying to sort this- they stated it would be paid out in 48 hours- still waiting!!!! TO EVERYBODY WHO HAS AN ACCOUNT MOVE YOU'RE MONEY AS SOON AS YOU CAN- THIS BANK IS GOING BROKE!!!!!!	1
3	santander	Anna	46	Lovely digital front , when you hit submit it's all paper	Lovely digital front , when you hit submit it's all paperThey have a nice front end digital application, as soon as you hit submit it's all very manual and paper based. Very old school for a bank that survived the pandemic	1
4	santander	Artem Chubarov	10	Unacceptably long waiting times	Unacceptably long waiting timesI spent 3(!!!) hours on Saturday during my weekend, waiting for the help desk t o reply and assist me with a blocked transfer. I was buying a car from a dealer and used a bank transfer to pay. However, the transfer was blocked by Santander and I got stuck for 4+ hours at the dealership office (which wa s supposed to be closed already back then), waiting for the help desk to unfreeze my transaction. It took me 1 hour of waiting to hear the actual person for the first time, who then put me in another queue for another hour . Needless to say, there were no updates while I was waiting - just the background music.Whatever the reason for the stuff shortage is, that service level is horrendous and a bsolutely unbearable. I've been with Santander for 5+ years but definitely switching away as soon as possible.	1

Figure 2 Scrapped Dataset('banks.data') examples

banks.data[[1]]; Traditional Banks

Bank\Score	1	2	3	4	5	Total	Percentage
Barclays	5,247	181	80	140	545	6,193	28%
HSBC	5,077	146	79	72	333	5,707	26%
Nationwide	2,041	160	62	93	446	2,802	13%
Natwest	3,026	130	56	66	312	3,590	16%
Santander	3,440	162	38	93	331	4,064	18%
Total	18,831	779	315	464	1,967	22,356	100%
Percentage	84%	3%	1%	2%	9%	100%	

banks.data[[2]]; Digital Banks

Bank\Score	1	2	3	4	5	Total	Percentage
monzo	2,060	251	330	1,297	16,325	20,263	15%
revolut	6,680	1,208	2,251	9,323	69,414	88,876	65%
starling	2,625	454	742	2,834	20,584	27,239	20%
Total	11,365	1,913	3,323	13,454	106,323	136,378	100%
Percentage	8%	1%	2%	10%	78%	100%	

Figure 3 'banks.data' composition ratio

After scrapping, seven data cleaning steps are done to improve the data quality. First, id, and other columns for further analysis are created. Then, reviews written in English and users with at least 3 reviews are kept by using cus_tot_review column. Next, text cleaning, tokenisation, removing stop-words, lemmatisation, and spelling check are performed. The results are saved into a list named 'banks.clean'. Figure 4 and 5 illustrate examples and the composition ratio of the 'banks.clean'.

id	company	cus_name	cus_tot_review	title	review.only	review.total	score	cleanText
2	monzo	Anthony Cox	3	Banking as it should be	Amazing app, features and quick help when you need. Analytics, saving pots and roundups, leaves high street banks for dead.	Banking as it should be Amazing app, features and quick help when you need. Analytics, saving pots and roundups, leaves high street banks for dead.	5	banking as it should be amazing app features and quick help when you need analytics saving pots and roundups leaves high street banks for dead
3	monzo	Paul Savage	7	No need for a branch	Enjoying the experience of using this bank. Haven't missed branches. Not tested as no issues. Happy so far.	No need for a branch Enjoying the experience of using this bank. Haven't missed branches. Not tested as no issues. Happy so far.	5	no need for a branch enjoying the experience of using this bank havent missed branches not tested as no issues happy so far
8	monzo	Ex-customer	4	A good bank	Monzo have been fine. Personally got no complaints. Love the savings pots. Would like them to start making money now so then I know my money is safe.	A good bank Monzo have been fine. Personally got no complaints. Love the savings pots. Would like them to start making money now so then I know my money is safe.	4	a good bank monzo have been fine personally got no complaints love the savings pots would like them to start making money now so then i know my money is safe

Figure 4 Cleaned Dataset(banks.clean) examples

banks.clean[[1]]; Traditional Banks

Bank\Score	1	2	3	4	5	Total	Percentage
Barclays	3,230	137	57	114	368	3,906	28%
HSBC	2,998	107	56	51	212	3,424	24%
Nationwide	1,411	123	50	81	353	2,018	14%
Natwest	1,810	92	51	57	211	2,221	16%
Santander	2,118	121	25	77	243	2,584	18%
Total	11,567	580	239	380	1,387	14,153	100%
Percentage	82%	4%	2%	3%	10%	100%	

banks.clean[[2]]; Digital Banks

Bank\Score	1	2	3	4	5	Total	Percentage
monzo	931	127	149	469	6,311	7,987	19%
revolut	2,642	384	524	1,747	13,506	18,803	46%
starling	1,345	252	394	1,463	10,784	14,238	35%
Total	4,918	763	1,067	3,679	30,601	41,028	100%
Percentage	12%	2%	3%	9%	75%	100%	

Figure 5 'banks.clean' composition ratio

With cleaned dataset, BoW(Bag-of-Words) model and DTM(Document-Term-Matrix) of unigram, bigram and trigram are used to create features. Each review is regarded as a document for BoW Model and DTM and reviews with the same bank type and score are regarded as a document for TF-IDF. Figure 6 shows the words with highest TF-IDF value for each bank type and score. Most important words of both bank types are related to customer service. Other than that, customers are not satisfied with traditional banks because of sponsoring corrupt premier league and waiting time, whereas satisfied with the fact that banks are government verified banks. In contrast, customers are satisfied with digital banks' mobile applications and straightforward processes.

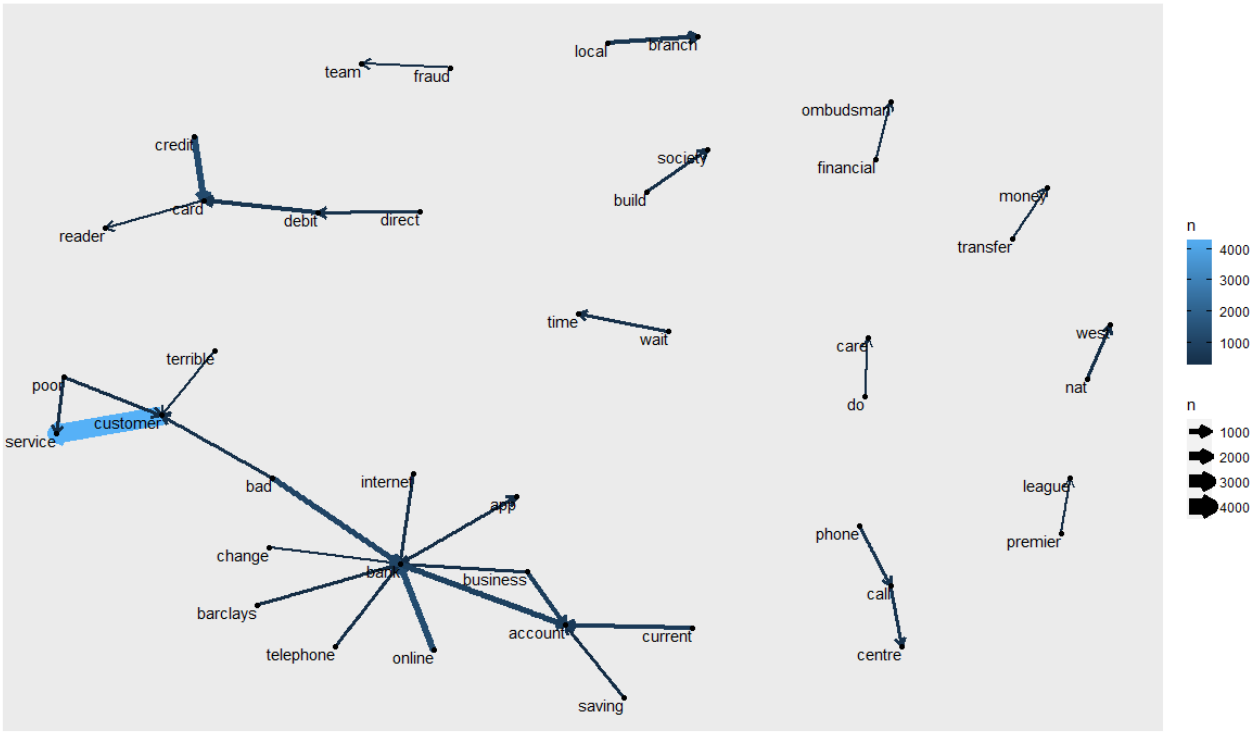
From word network graphs [Figure 7], which illustrate how the words are related to each other, 'customer service' is the most notable word combination from both types (Zhang, 2022). 'highly recommend' and 'exchange rate' show strong relations in digital banks, while 'bad bank' and 'local branch' are significant from traditional banks

Figure 8 shows the sparsity, the proportion of sparse(zero) entries in the entire DTM, increases as the number of words consisting of unit tokens increases. All sparsity values in Figure 8 are reasonable for

further analysis. Lastly, to explore the relationship between mentioning the bank name on the review with the score, a binary variable is created, and regression is done on Part B.

Figure 6 TF-IDF Model results

Traditional Banks



Digital Banks

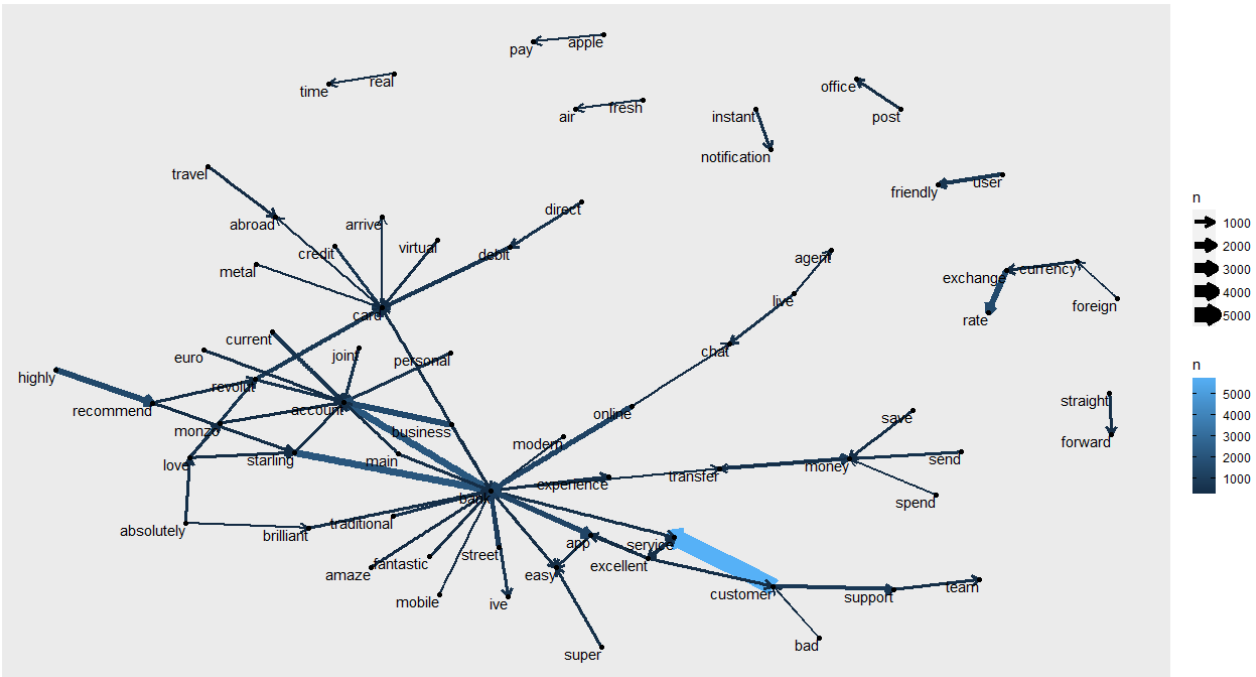


Figure 7 Bigram word networks

	Traditional Banks	Digital Banks
Unigram	75%	72%
Bigram	79%	77%
Trigram	80%	79%

Figure 8 DTM sparsity of each dataset

Part B

The aim of sentiment analysis in this report is to estimate the predictability of the review score. Sentiment analysis consists of three steps: calculating dictionary coverage, sentiment score, and regression for feature selection [Figure 9]. The sentiment analysis framework [Figure 10] illustrates how various sentiment analysis methods are adopted.

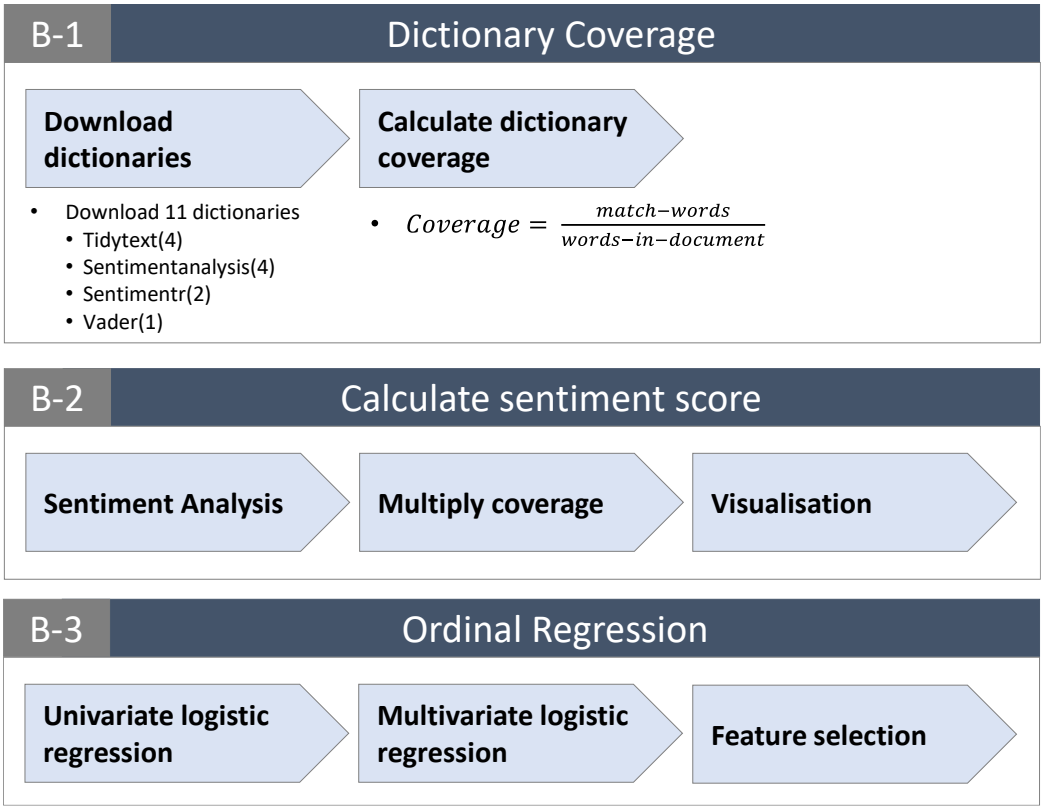


Figure 9 Part B Workflow

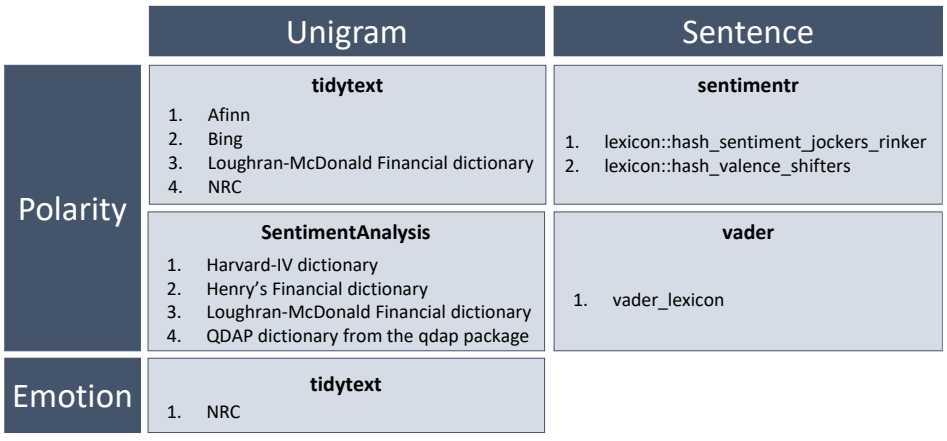


Figure 10 Sentiment Analysis Framework

For unigram sentiment analysis, eight dictionaries are matched with the tokenised words from the reviews. Since every dictionary has different word lists, dictionary coverage is calculated as coverage = match-words/words-in-documents, and the sentiment score is multiplied by coverage (Liske, 2018). Figures 11 and 12 show polarity differences between reviews with 1 and 5 score reviews from tidytext and SentimentAnalysis packages.

Some dictionaries reflect the sentiment well for the cases, but it is hard to find a dictionary that reflects the sentiment well for both bank types and scores. For example, polarity of Loughran dictionary of digital banks' score 1 are mostly negative, which means it reflects the sentiments well. However, it also shows a strong negative result for digital banks' score 5, contradicting the common sense that high score review are written with positive sentiment.

The NRC dictionary gives 10 emotion categories except for compound sentiment score, and Figure 13 show the correlation between emotions. Although opposite emotions like disgust and joy are positively related, the correlation between emotions and compound sentiment score(last variable) is reasonable.

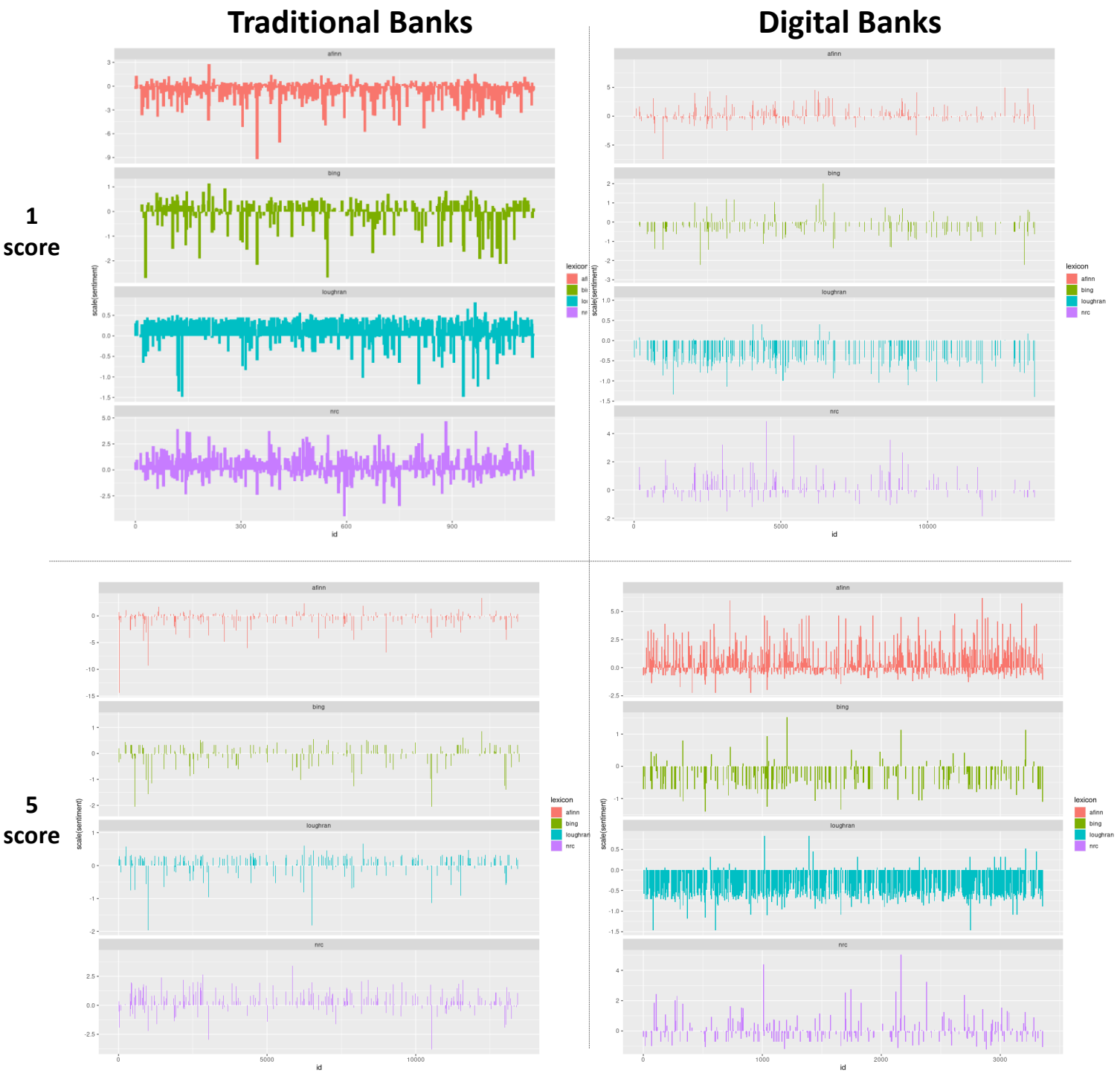


Figure 11 Polarity - tidytext package 4 dictionaries (Afinn(red), Bing(green), Loughran(blue), NRC(purple))

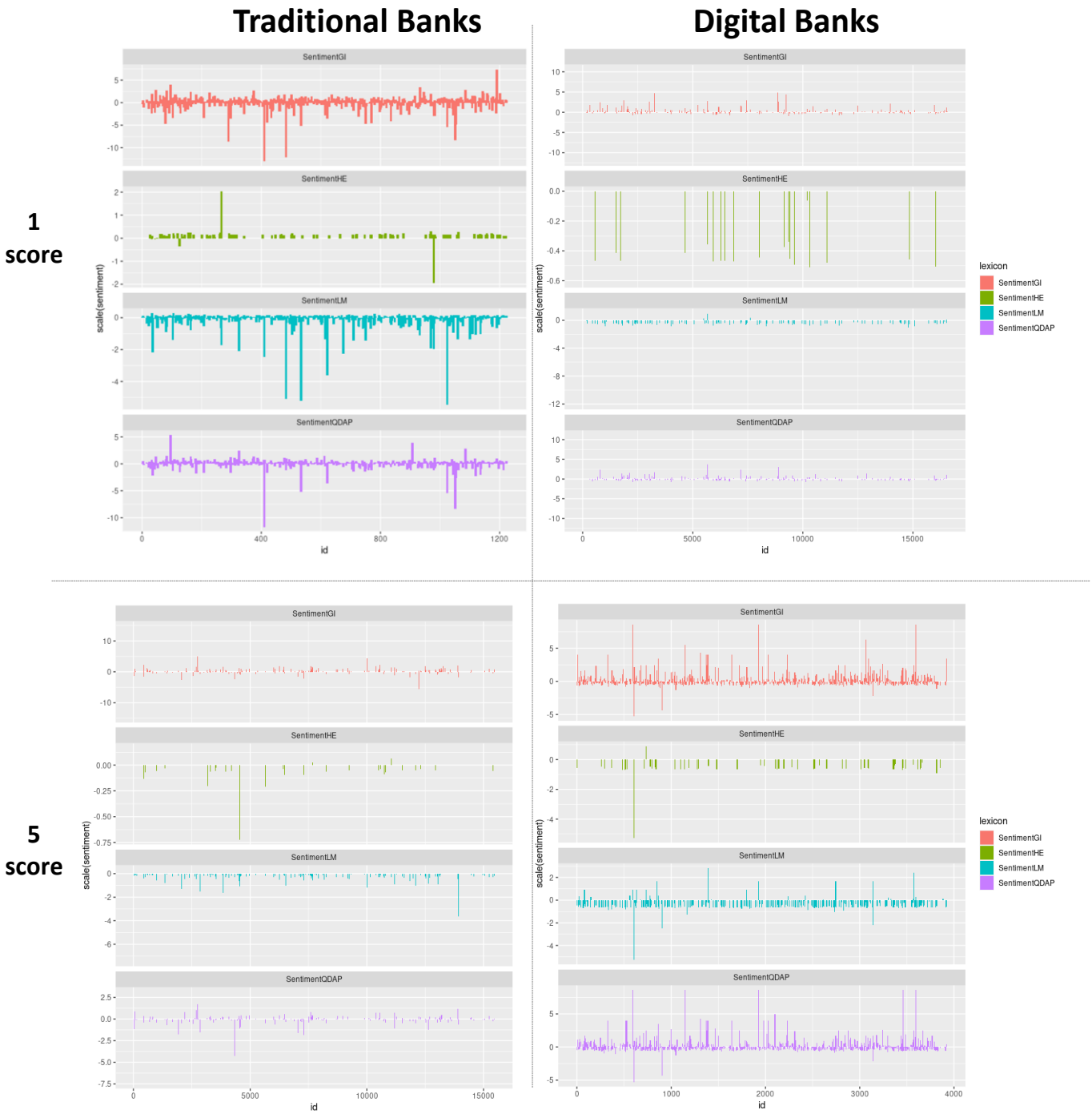


Figure 12 Polarity - SentimentAnalysis package 4 dictionaries (GI(red), HE(green), LM(blue), QDAP(purple))

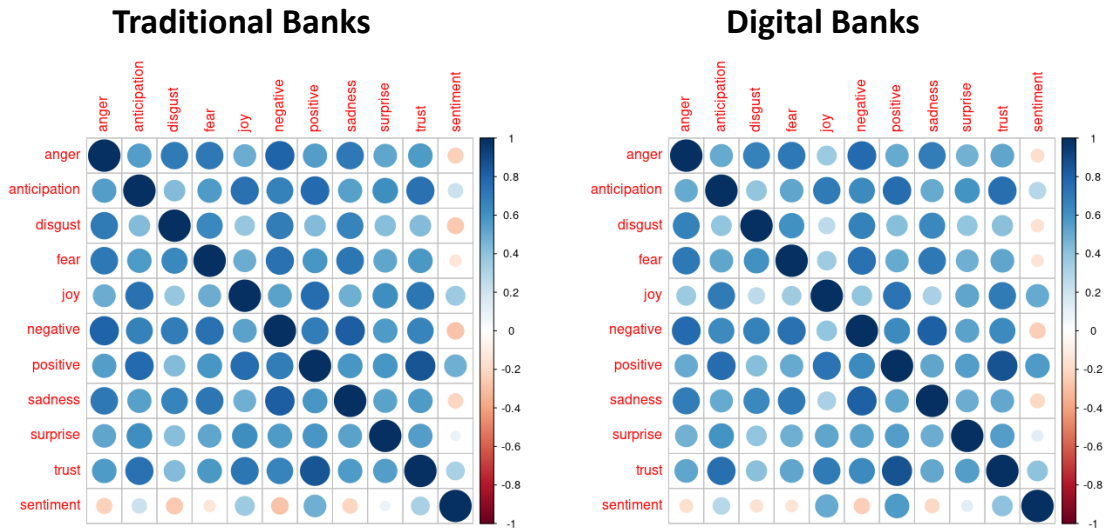


Figure 13 NRC Emotion Correlation

The unigram sentiment analysis provides some insight, but it cannot consider the context, such as valence shifters. Therefore, sentiment analysis on sentences is performed using sentimentr and vader(Valence-Aware-Dictionary-for-Sentiment-Reasoning) packages. 'vader' can deal with polarity and intensity by taking into account capital letter, amplifiers (e.g., very), modifiers (e.g. slightly), negators (e.g., not) and some punctuation signs(e.g., !) (Garcia, 2021). 'sentimentr' also can deal with valence shifters while maintaining speed (RinkerTyler, 2022). Figures 14 and 15 illustrate the polarity difference between reviews with 1 and 5 score reviews from sentiment and vader packages, respectively. The average of 5 score reviews' sentiment score is higher than that of 1 score review's sentiment score for both packages. Readability, the number of words written in capitals and exclamation marks are calculated as additional features to estimate the review score.

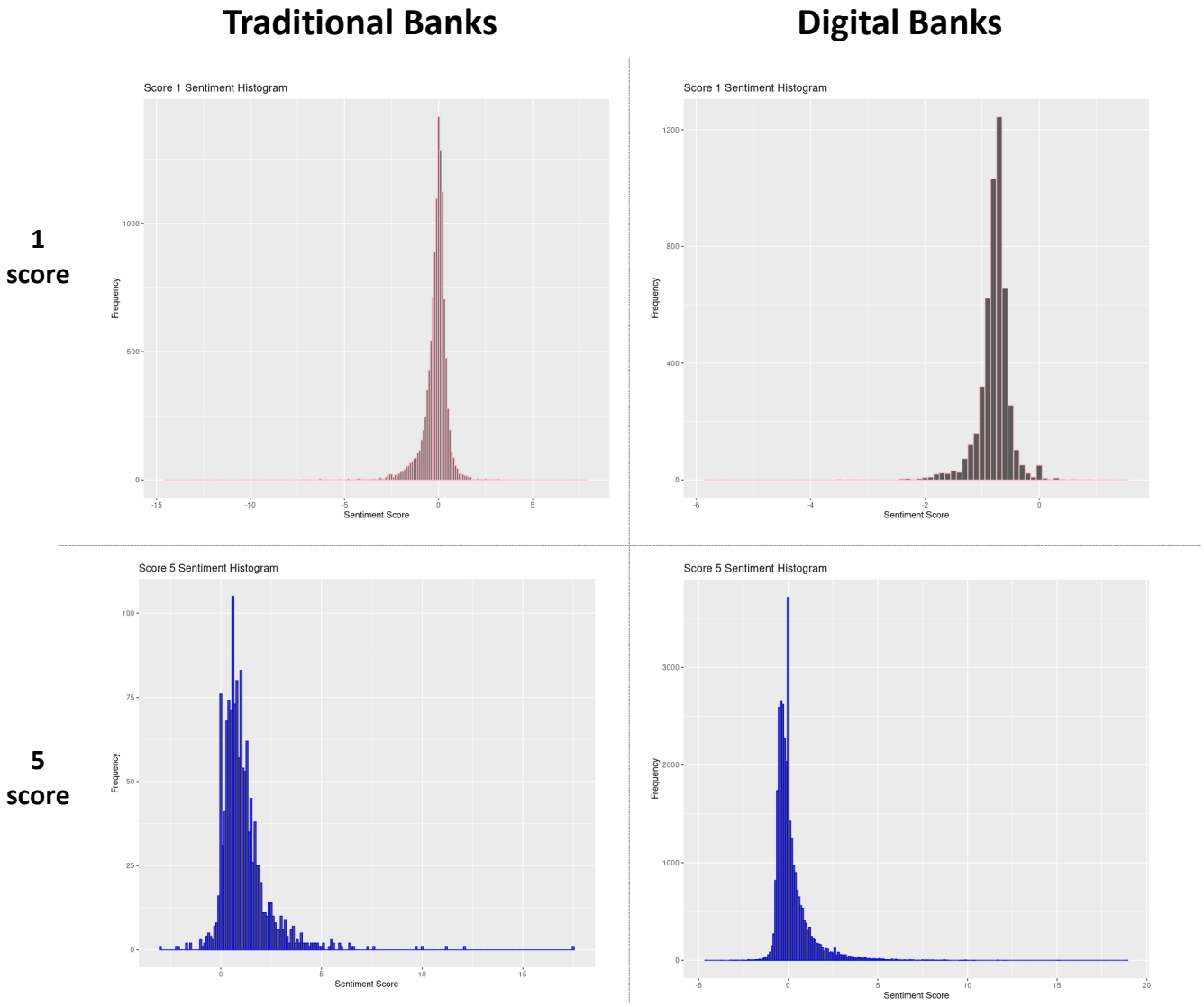


Figure 14 Polarity - Sentimentr package, the average of 5 score review is higher than the average of 1 score review

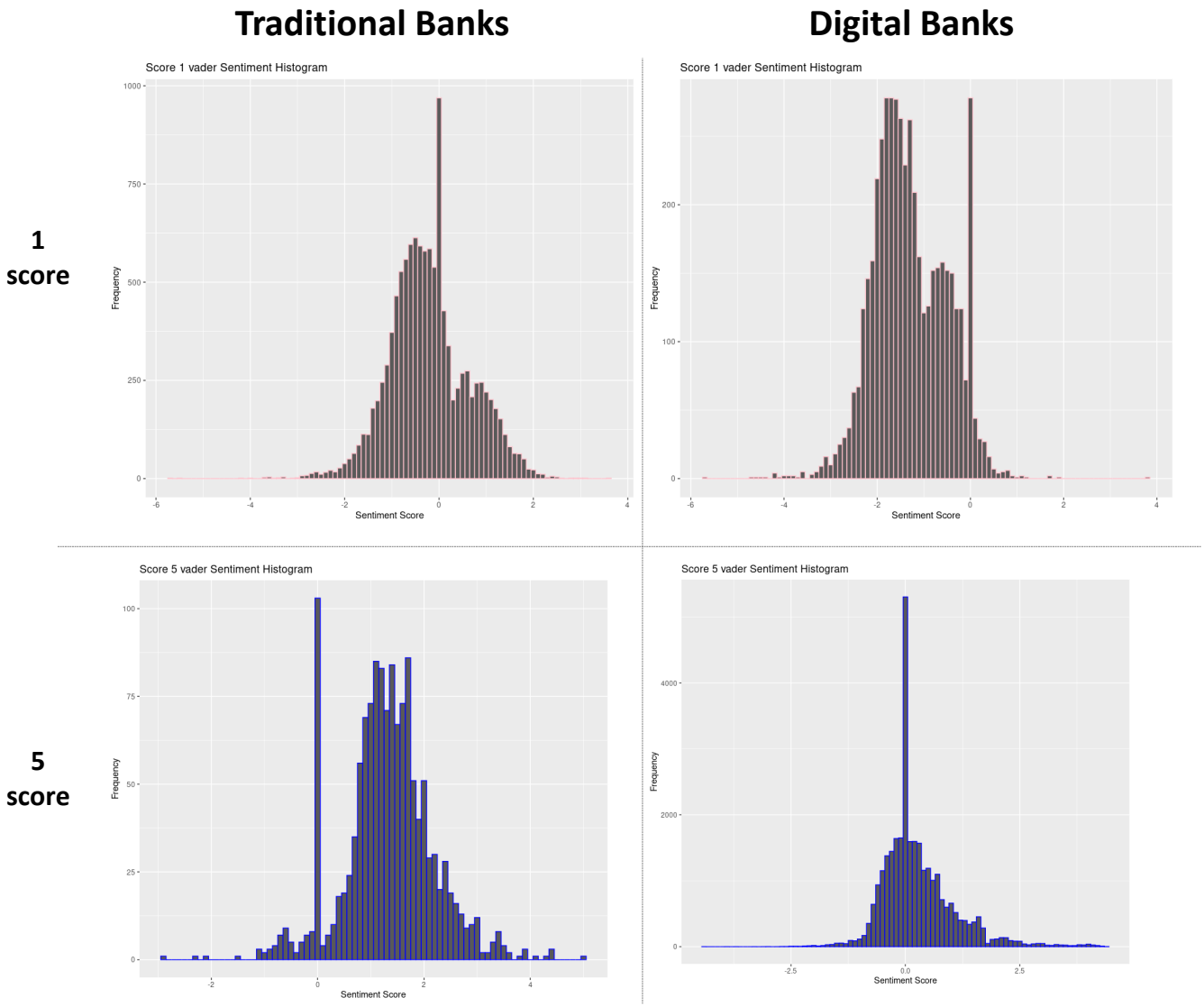


Figure 15 Polarity – vader package, the average of 5 score review is higher than the average of 1 score review

Finally, feature selection for predicting score is done using logistic regression. All sentiment scores and other features are significant in univariate ordinal regression. Therefore, clockwise multivariate ordinal regression is used and the results are shown in Figure 16 (How to Perform Ordinal Logistic Regression in R | R-bloggers, 2022). From the multivariate regression result, insignificant features are removed. Any sentiment score with negative relation to score is removed even though its p-value is significant because it contradicts the common sense that positive sentiment is related to a high score, which might be related to skewed dataset. The number of words written in capitals and exclamation marks are statistically significant but also removed because vader score considers them, and it is hard to derive more value from them. Some emotion scores from NRC dictionaries are statistically significant, but because of low coverage(7% for digital banks), they are removed as well [Figure 17]. Features with coefficients higher than 0.5 are selected among statistically significant sentiment features. Figure 18 shows that sentimenr score and vader score are selected among sentiment features, and 'bank_mentioned' and 'readability' are selected as the Part B final models.

	Dependent variable:														
	score														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
sentiment_bing	1.284***														0.094*
	(0.044)														(0.055)
sentiment_ld		1.295***													0.183***
		(0.044)													(0.050)
sentiment_afinn			1.660***												0.284***
			(0.041)												(0.052)
sentiment_nrc				0.510***											0.038
				(0.024)											(0.036)
SentimentGI					0.640***										-0.143***
					(0.029)										(0.041)
SentimentHE						0.279***									-0.105
						(0.093)									(0.066)
SentimentLM							0.790***								-0.165***
							(0.061)								(0.051)
SentimentQDAP								1.201***							0.205***
								(0.042)							(0.052)
ave_sentiment									1.818***						0.867***
									(0.042)						(0.046)
vader										1.489***					0.953***
										(0.029)					(0.040)
capital_scale											-6.535***				-3.812***
											(0.905)				(0.805)
exclaim_scale												-0.516***			-0.467***
												(0.076)			(0.084)
bank_mentioned													0.059***		0.167***
													(0.022)		(0.027)
Automated_Readability_Index														-0.332***	-0.316***
														(0.030)	(0.040)
Observations	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153	14,153
Note:															*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:														
	ordered logistic							score cumulative link					ordered logistic		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
sentiment_bing	2.133***														0.310***
	(0.036)														(0.037)
sentiment_ld		2.051***													0.451***
		(0.026)													(0.032)
sentiment_afinn			2.171***												0.345***
			(0.027)												(0.033)
sentiment_nrc				0.809***											0.242***
				(0.023)											(0.024)
SentimentGI					1.336***										-0.152***
					(0.027)										(0.025)
SentimentHE						0.989***									-0.185***
						(0.096)									(0.056)
SentimentLM							3.876***								0.258***
							(0.077)								(0.038)
SentimentQDAP								2.924***							0.075**
								(0.043)							(0.036)
ave_sentiment									2.794***						0.857***
									(0.036)						(0.037)
vader										2.015***					1.090***
										(0.020)					(0.027)
capital_scale											-7.661***				-3.676***
											(0.468)				(0.314)
exclaim_scale												0.305***			0.146***
												(0.022)			(0.025)
bank_mentioned													-0.226***		0.264***
													(0.011)		(0.015)
Automated_Readability_Index														-1.085***	-0.693***
														(0.015)	(0.019)
Observations	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028	41,028
Log Likelihood							-33,116.550	-31,552.850	-29,804.210						
Note:															*p<0.1; **p<0.05; ***p<0.01

Figure 16 Sentiment Analysis regression reslt before feature selection – Traditional banks(top), Digital Banks(bottom)

	Dependent variable:	
	score	
	(1)	(2)
anger	-0.036	-0.122***
	(0.045)	(0.040)
anticipation	-0.090**	-0.063**
	(0.036)	(0.028)
disgust	-0.264***	-0.274***
	(0.062)	(0.055)
fear	-0.015	0.028
	(0.042)	(0.040)
joy	0.268***	0.481***
	(0.049)	(0.038)
negative	-0.158***	-0.220***
	(0.033)	(0.027)
positive	0.117***	0.054***
	(0.023)	(0.019)
sadness	0.061	-0.033
	(0.048)	(0.044)
surprise	0.041	-0.188***
	(0.050)	(0.044)
trust	-0.020	-0.039*
	(0.027)	(0.021)
Observations	4,300	2,847
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 17 NRC emotion regression result - Traditional banks(1), Digital Banks(2)

	Dependent variable:					
	score					
	(1)	(2)	(3)	(4)	(5)	(6)
sentiment_bing	0.094*	0.140***		0.310***	0.502***	
	(0.055)	(0.051)		(0.037)	(0.037)	
sentiment_ld	0.183***	0.230***		0.451***	0.786***	0.724***
	(0.050)	(0.047)		(0.032)	(0.031)	(0.028)
sentiment_afinn	0.284***	0.317***		0.345***	0.314***	
	(0.052)	(0.051)		(0.033)	(0.033)	
sentiment_nrc	0.038			0.242***	-0.112***	
	(0.036)			(0.024)	(0.021)	
SentimentGI	-0.143***			-0.152***		
	(0.041)			(0.025)		
SentimentHE	-0.105			-0.185***		
	(0.066)			(0.056)		
SentimentLM	-0.165***			0.258***	0.083*	
	(0.051)			(0.038)	(0.046)	
SentimentQDAP	0.205***	0.057		0.075**	0.186***	
	(0.052)	(0.045)		(0.036)	(0.039)	
ave_sentiment	0.867***	0.913***	0.993***	0.857***	0.978***	0.945***
	(0.046)	(0.046)	(0.044)	(0.037)	(0.037)	(0.037)
vader	0.953***	0.881***	1.108***	1.090***	1.118***	1.305***
	(0.040)	(0.038)	(0.034)	(0.027)	(0.027)	(0.024)
capital_scale	-3.812***			-3.676***		
	(0.805)			(0.314)		
exclaim_scale	-0.467***			0.146***		
	(0.084)			(0.025)		
bank_mentioned	0.167***	0.070***	0.161***	0.264***	0.090***	0.280***
	(0.027)	(0.026)	(0.027)	(0.015)	(0.014)	(0.015)
Automated_Readability_Index	-0.316***		-0.370***	-0.693***		-0.611***
	(0.040)		(0.035)	(0.019)		(0.017)
Observations	14,153	14,153	14,153	41,028	41,028	41,028
Note:						*p<0.1; **p<0.05; ***p<0.01

Figure 18 Part B Final model - Traditional banks(3), Digital banks(6)

Part C

In Part C, topic models are built using STM(Structural-Topic-Model) and sLDA(Supervised Latent-Dirichlet Allocation-model) and used to estimate the predictability of the score [Figure 19]. First, tokenisation is done using Part-of-speech tagging, and then nouns, adjectives, and adverbs are selected. Next, tokens are cleaned by removing punctuations, stopwords and numbers, then filtered, which appear at least 1% of the time in the document corpus.

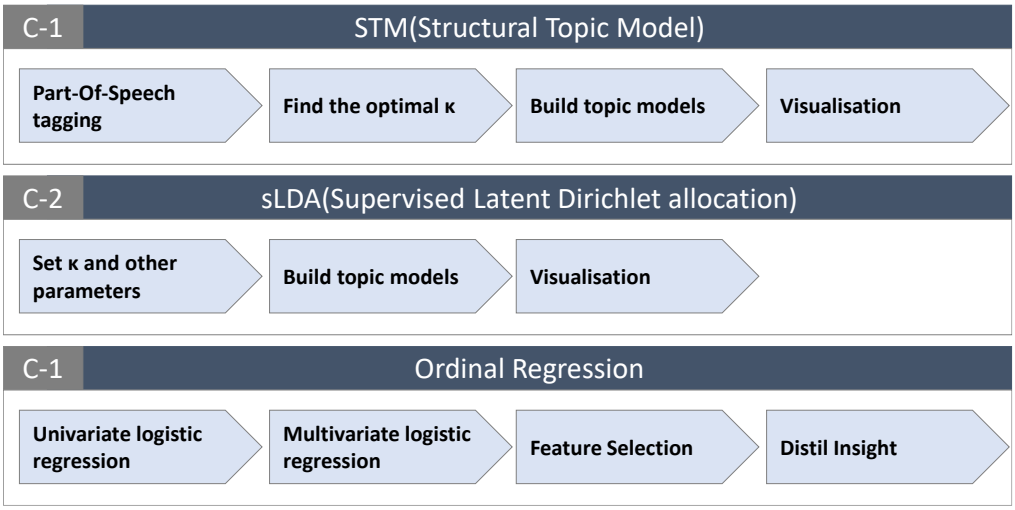


Figure 19 Part C Workflow

Then, based on Held-out likehood, residuals, semantic coherence and lower bound, κ is selected as 14 [Figure 20]. Perplexity is optimised when using STM by default (Package 'stm', 2020). According to graphs, a higher value of κ could be considered, but the number of topics is limited for interpretation. After building models, each topic is labelled using the most representative documents using findThoughts function, word cloud [Appendix Figure A-27:30] and Frex(Frequency and Exclusivity) words [Appendix Figure-36] of each topic (findThoughts function, 2022).

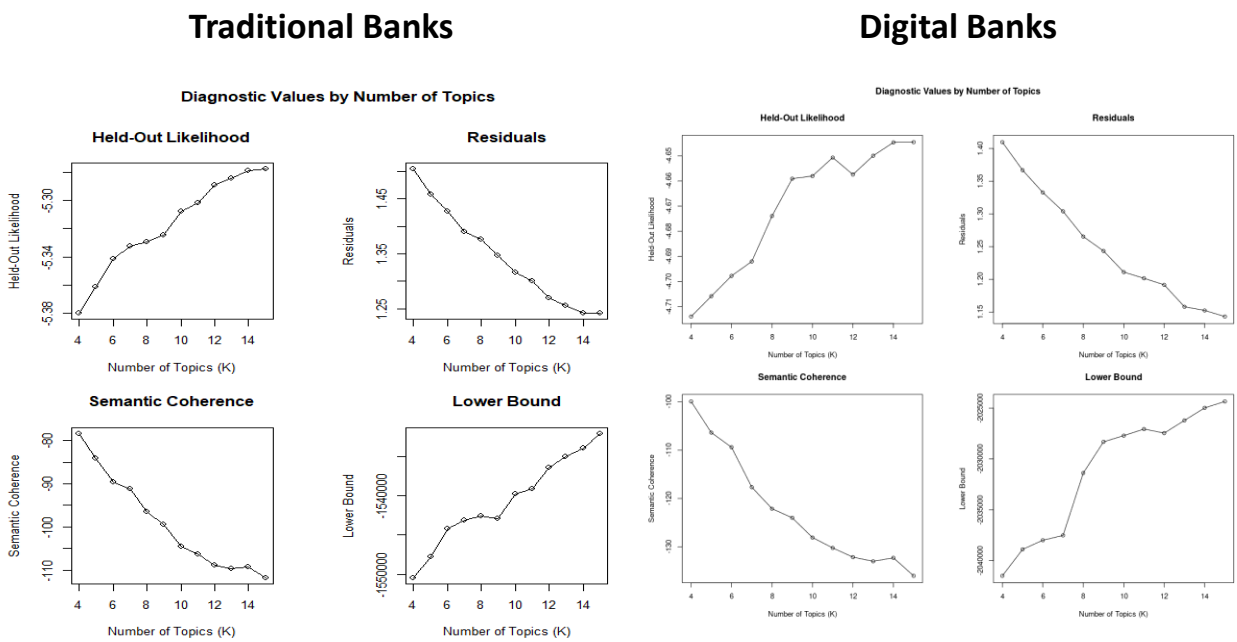


Figure 20 Diagnostic values by the number of topics

Then marginal effects graphs of each topic against the score are plotted in Figure 21. The graphs illustrate that customers are satisfied with branch banking while highly dissatisfied with the long waiting time. In contrast, customers of digital banks like user-friendly applications and simple online banking but are not satisfied with issues on money-back and business accounts. The individual marginal effects [Appendix Figure A-31:35] show that "satisfaction on branch banking" has the highest expected topic proportion in traditional banks, and "issue on money-back" has the highest expected topic proportion in digital banks.

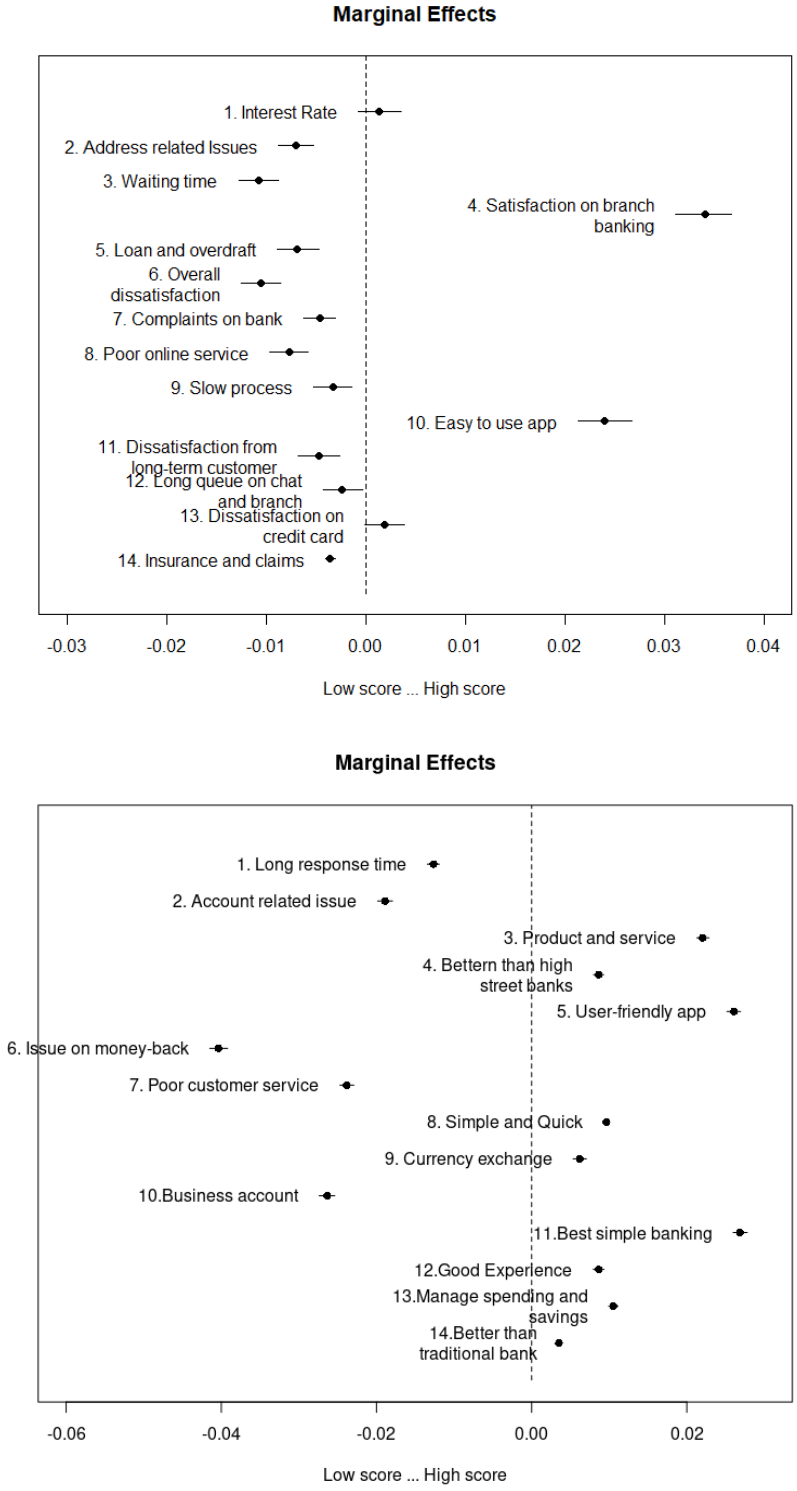


Figure 21 Marginal Effects - Traditional Banks(top), Digital Banks(bottom)

Usually, an unsupervised learning approach is adopted for topic modelling to "find the structure between topics and terms and the relationship between document and topics" (Zhang, 2022). sLDA, a supervised approach, enables to prediction of an observed univariate outcome setting the parameters extracted from the unsupervised learning approach, and simultaneously fitting a topic model (Zhang, 2022). The optimal topic number κ from the previous step(STM) is used, and the results of sLDA predicting score are shown in Figure 22. Most topics of traditional banks estimate low scores, while topics related to in-person and mortgage estimate high scores. In contrast, most topics on digital banks predict high scores, while topics related to a business account and crypto estimate low scores. By analysing the STM and sLDA, it is recommended that traditional banks improve customer service in terms of response time and the communication channel (paperless banking) and keep focusing on corporate banking and loans, especially mortgages.

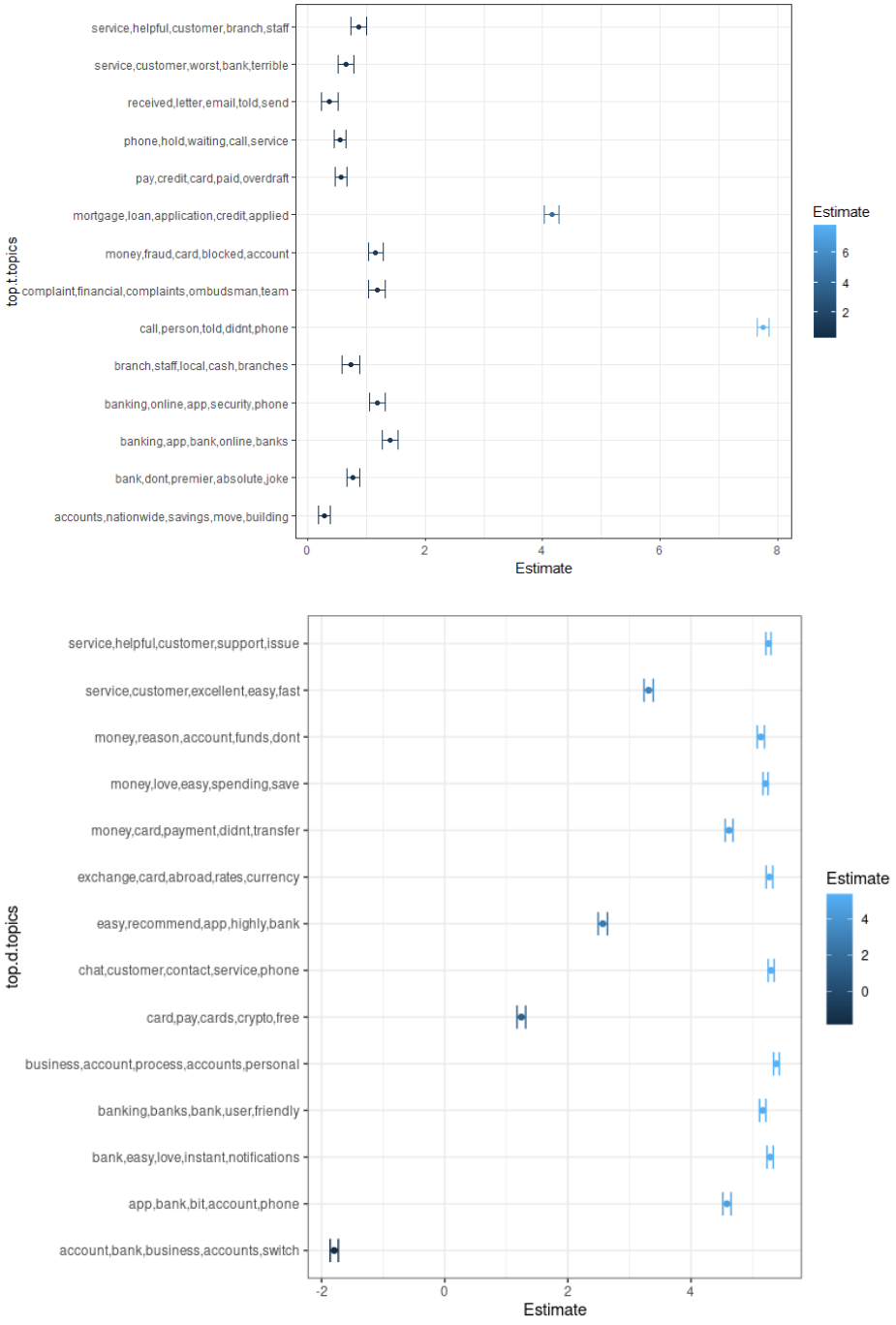


Figure 22 sLDA result Traditional Banks(top), Digital Banks(bottom)

Lastly, topics are used to fit logistic regression models by combining the significant features from Part B. In Figure 23, the final model of traditional banks has 16 variables; 12 topics, 2 sentiment features and 2 other text features. The final model of digital banks has 7 topics, 1 sentiment feature, and 2 other text features. Since the dependent variable is ordinal, it is not clear to capture the variance of the model or compare the goodness-of-fit. However, it is still tempted to compare the models using R^2, therefore the Hosmer and Lemeshow suggest "...low R^2 values in logistic regression are the norm and this presents a problem when reporting their values to an audience accustomed to seeing linear regression values. ... Thus we do not recommend routine publishing of R^2 values with results from fitted logistic models. However, they may be helpful in the model building state as a statistic to evaluate competing models." (Hosmer & Lemeshow, 2010). Figure 24 suggests that model fit is improved by considering topics for both bank types compared to part B's final model.

	Dependent variable:		
	score		
	(1)	(2)	(3)
topic_1	13.363***	8.148***	6.716***
	(1.208)	(1.186)	(0.578)
topic_2	10.077***	6.531***	5.054***
	(1.284)	(1.249)	(0.650)
topic_3	7.315***	4.120***	2.688***
	(1.249)	(1.219)	(0.644)
topic_4	17.935***	12.961***	11.549***
	(1.170)	(1.148)	(0.525)
topic_5	9.734***	4.885***	3.421***
	(1.252)	(1.229)	(0.626)
topic_6	4.023***	1.838	
	(1.333)	(1.313)	
topic_7	11.196***	6.756***	5.223***
	(1.351)	(1.329)	(0.736)
topic_8	9.045***	5.042***	3.636***
	(1.213)	(1.179)	(0.597)
topic_9	11.264***	6.952***	5.494***
	(1.246)	(1.221)	(0.617)
topic_10	16.745***	11.493***	10.093***
	(1.160)	(1.136)	(0.515)
topic_11	9.846***	5.571***	4.130***
	(1.227)	(1.210)	(0.615)
topic_12	11.282***	6.859***	5.433***
	(1.211)	(1.183)	(0.580)
topic_13	13.469***	8.013***	6.581***
	(1.202)	(1.176)	(0.559)
ave_sentiment		0.749***	0.751***
		(0.047)	(0.047)
vader		0.884***	0.885***
		(0.036)	(0.036)
bank_mentioned		0.181***	0.180***
		(0.030)	(0.030)
Automated_Readability_Index		-0.129***	-0.137***
		(0.034)	(0.034)
Observations	14,085	14,085	14,085
Note:			*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:			
	score			
	(1)	(2)	(3)	(4)
topic_1	-3.553*			
	(2.034)			
topic_2	-2.860			
	(1.974)			
topic_3	35.339***	36.737***	34.953***	34.899***
	(2.285)	(1.010)	(0.990)	(0.989)
topic_4	3.123			
	(2.029)			
topic_5	3.771*			
	(2.001)			
topic_6	-13.987***	-16.117***	-15.264***	-15.245***
	(1.988)	(0.322)	(0.332)	(0.331)
topic_7	-6.998***	-8.119***	-7.636***	-7.618***
	(1.980)	(0.262)	(0.267)	(0.266)
topic_8	39.155***	39.587***	37.604***	37.508***
	(2.630)	(1.408)	(1.366)	(1.363)
topic_9	1.761			
	(1.980)			
topic_10	-8.215***	-10.150***	-9.306***	-9.291***
	(1.959)	(0.316)	(0.316)	(0.316)
topic_11	23.673***	24.885***	24.102***	24.058***
	(2.092)	(0.687)	(0.679)	(0.678)
topic_12	0.066			
	(2.033)			
topic_13	10.481***	10.073***	9.612***	9.617***
	(2.163)	(0.548)	(0.537)	(0.537)
topic_14		-3.449*		
		(1.850)		
ave_sentiment			-0.041	
			(0.033)	
vader			0.422***	0.409***
			(0.027)	(0.025)
bank_mentioned			0.208***	0.210***
			(0.019)	(0.019)
Automated_Readability_Index			-0.097***	-0.095***
			(0.017)	(0.017)
Observations	40,751	40,751	40,751	40,751
Note:			*p<0.1; **p<0.05; ***p<0.01	

Figure 23 Final model regression result - Traditional banks(left), Digital banks(right)

Nagelkerke's R2	Traditional Banks	Digital Banks
Part B Final model	0.468	0.342
Part C Final model	0.486	0.790

Figure 24 Final model R2 comparison

The final regression models suggest that it is essential for digital laggards to develop a user-friendly app with valuable features like instant notification on transactions, spending statistics, and budget management. Furthermore, reducing customer service response time based on fast, straightforward and digital-friendly processes is key to improving customer satisfaction. Additionally, focusing on corporate customers and loans can be a winning strategy for traditional banks.

For further study, consideration of the skewed dataset in bank types, review scores, and customer demographic features is recommended. For example, demographic features of the customer, such as age and sex, are not open to the public. However, a high proportion of the younger generation is expected to be keener to write online reviews than the older generation.

Bibliography

findThoughts function. (2022). Retrieved from Rdocumentation.org: https://www.rdocumentation.org/packages/stm/versions/1.3.6/topics/findThoughts

Garcia, D. D. (2021). *Running Unsupervised Sentiment Analysis in R.* Retrieved from https://dgarcia-eu.github.io/SocialDataScience/3_Affect/035_UnsupervisedToolsR/UnsupervisedToolsR.html

Hosmer, D., & Lemeshow, S. (2010). *Applied logistic regression.* New York: John Wiley: 167.

How to Perform Ordinal Logistic Regression in R | R-bloggers. (2022). Retrieved from R-bloggers: https://www.r-bloggers.com/2019/06/how-to-perform-ordinal-logistic-regression-in-r/

Lightico. (2022). *Consumer Study Details How Traditional Banks Should View Digital-Only Banks.* Retrieved from Lightico: https://www.lightico.com/blog/consumer-study-details-how-traditional-banks-should-view-digital-only-banks/

Liske, D. (2018, 03 29). *Tidy Sentiment Analysis in R.* Retrieved from Datacamp: https://www.datacamp.com/community/tutorials/sentiment-analysis-R

Package 'stm'. (2020). Retrieved from Cran.r-project.org: https://cran.r-project.org/web/packages/stm/stm.pdf

Rinker, T. (2022). *README*. Retrieved from Cran.r-project.org: https://cran.r-project.org/web/packages/sentimentr/readme/README.html#:~:text=sentimentr%20is%20designed%20to%20quickly,by%20the%20current%20R%20tools.

Selectorgadget.com. (2022). *SelectorGadget: point and click CSS selectors.* Retrieved from <https://selectorgadget.com/>

Statista. (2022). *Customers of leading UK banks 2021.* Retrieved from Statista: https://www.statista.com/statistics/940560/number-of-customers-at-select-banks-in-the-united-kingdom/

Trustpilot Reviews. (2022). Retrieved 04 09, 2022, from Trustpilot Reviews: https://uk.trustpilot.com/

Zhang, Z. (2022). *Text Mining for Social and Behavioral Research Using R.* Retrieved from Books.psychstat.org: https://books.psychstat.org/textmining/topic-models.html

Appendix

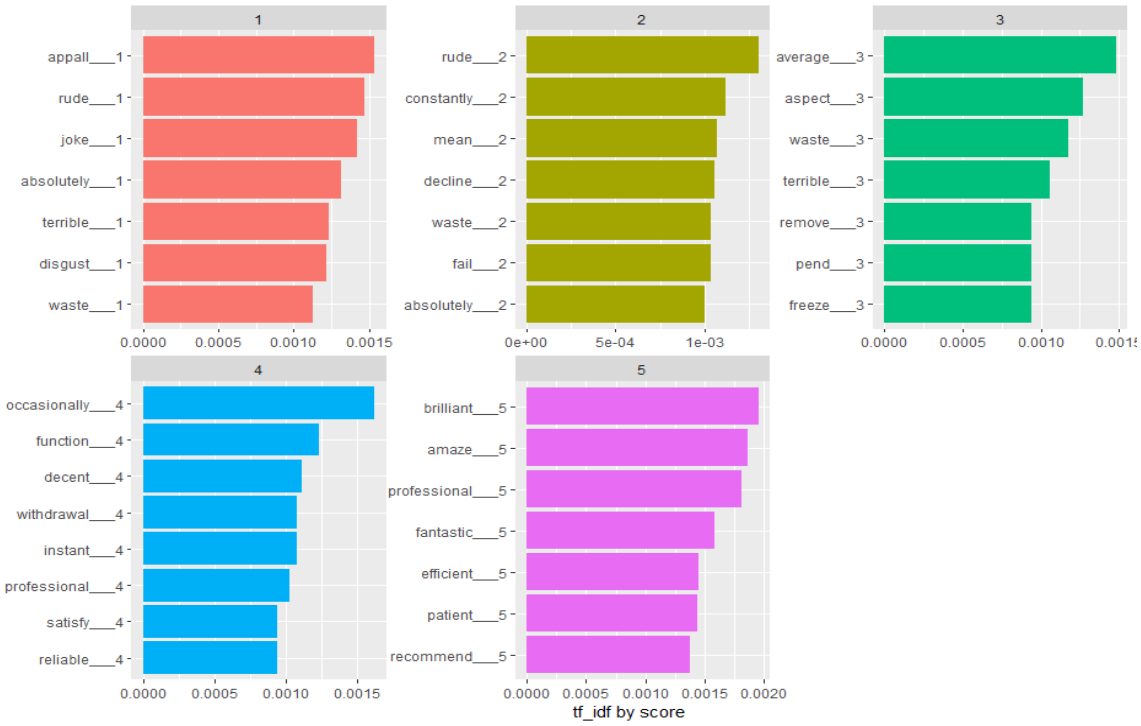


Figure A- 1 Traditional Banks unigram TF-IDF

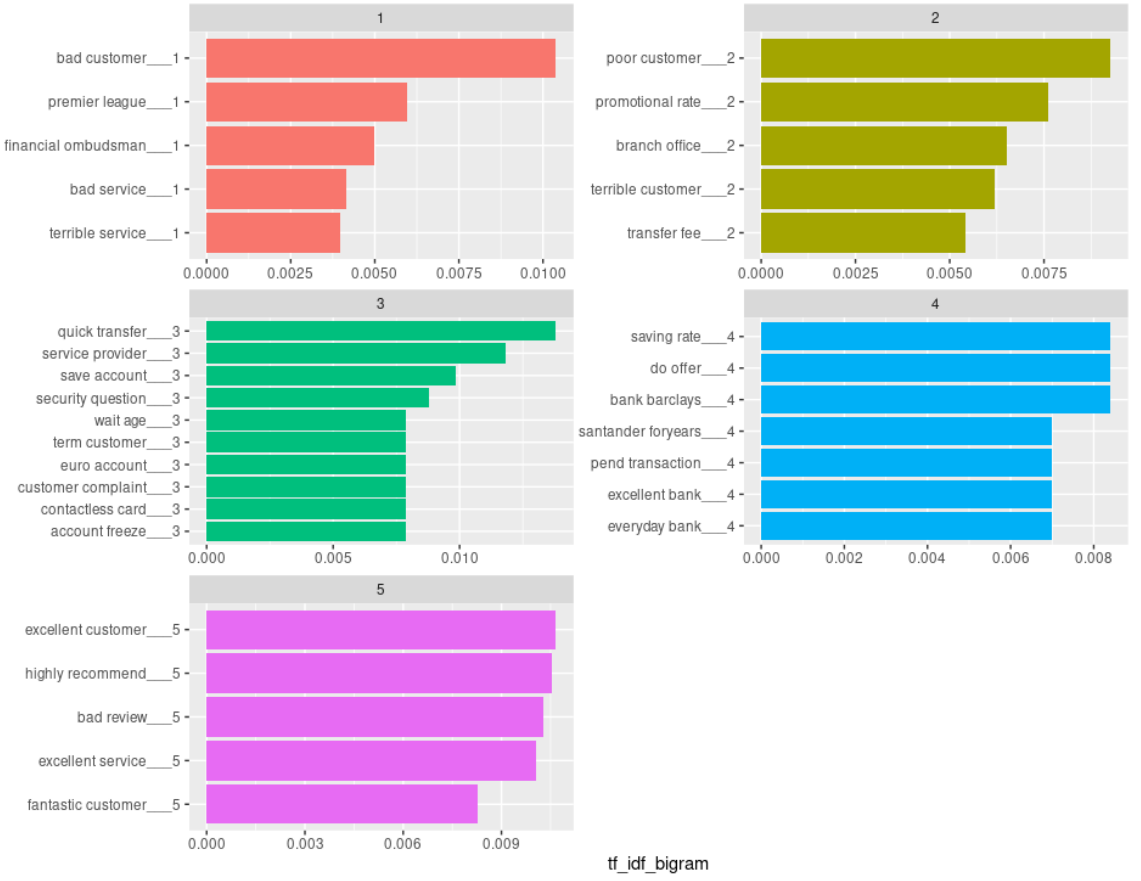


Figure A- 2 Traditional Banks bigram TF-IDF

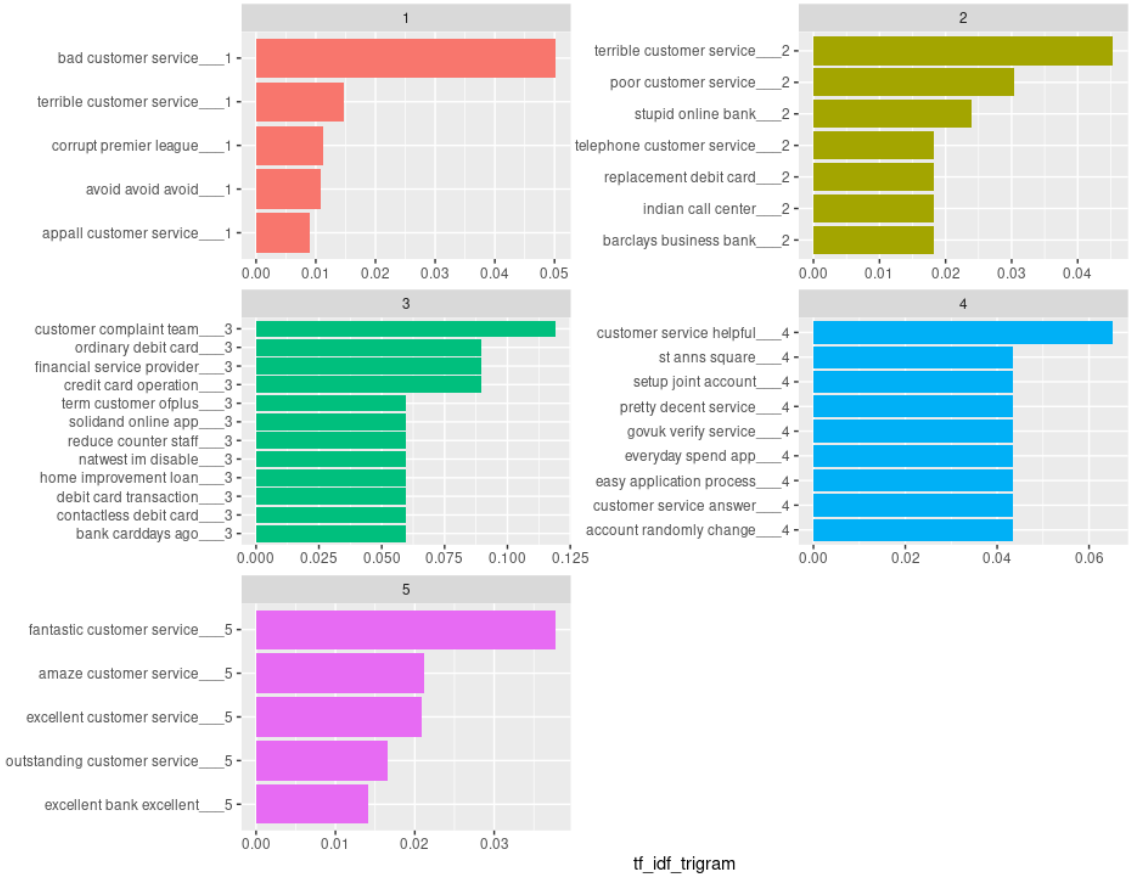


Figure A- 3 Traditional Banks trigram TF-IDF

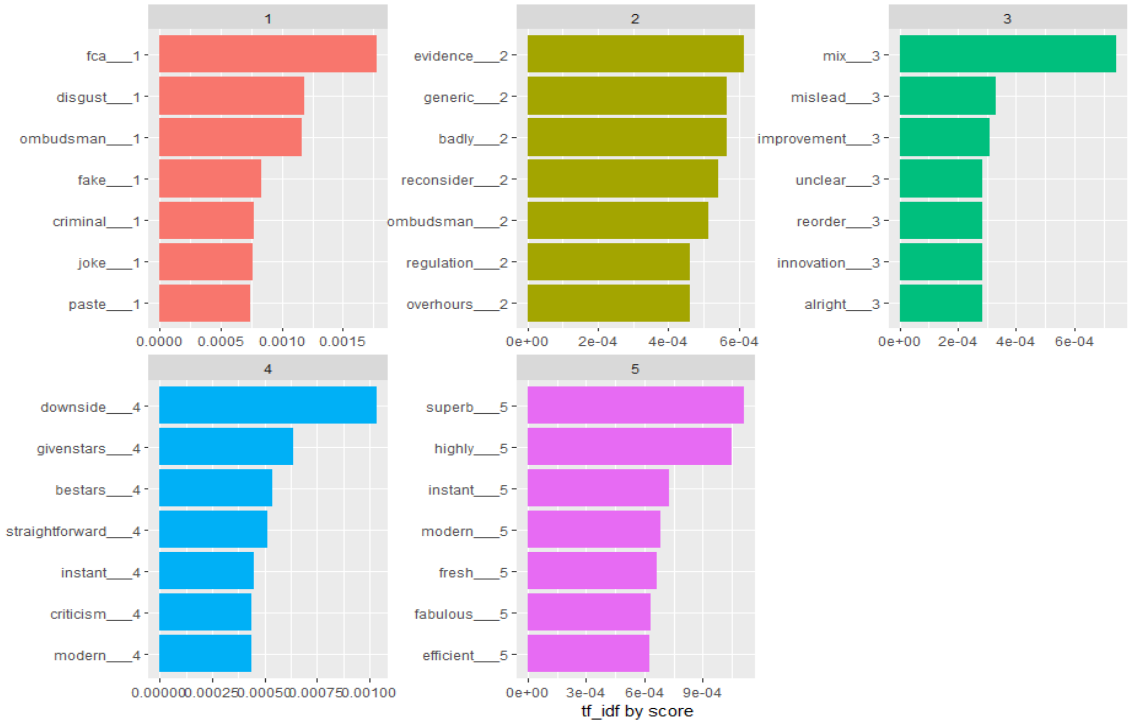


Figure A- 4 Digital Banks unigram TF-IDF

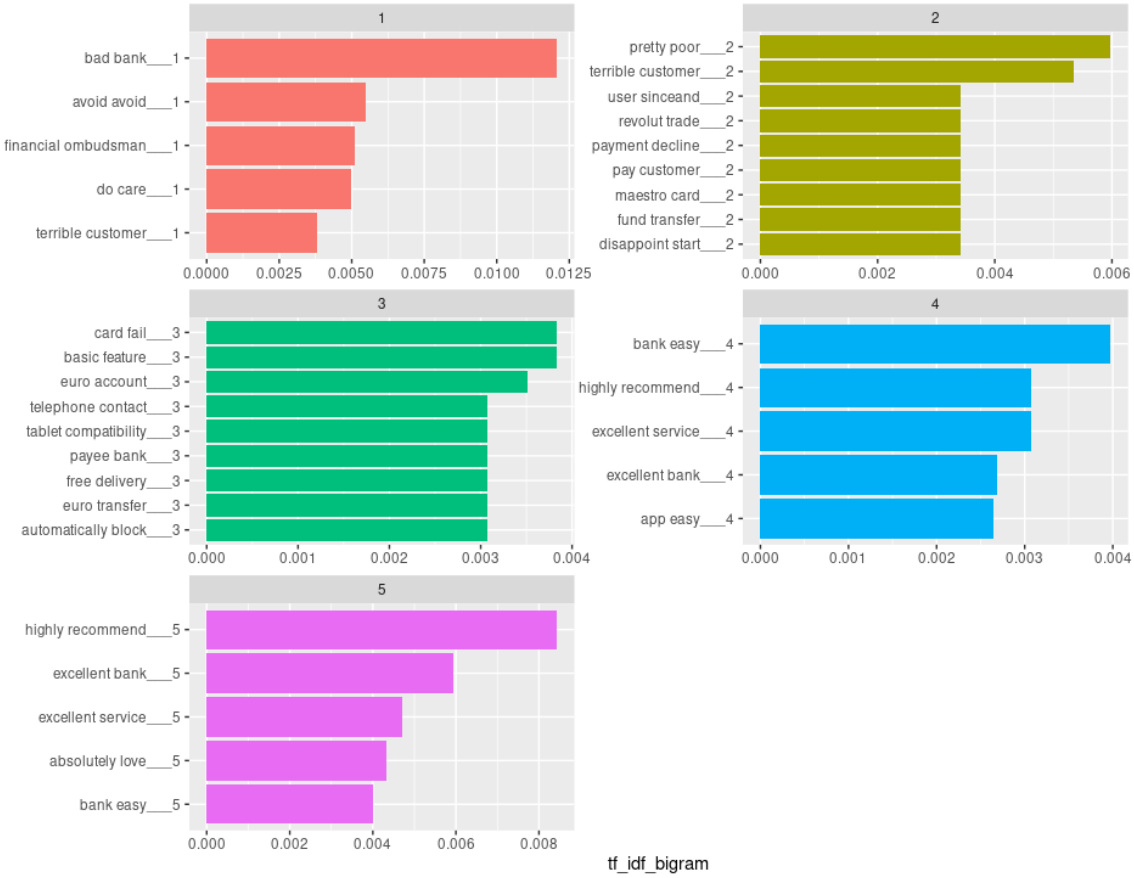


Figure A- 5 Digital Banks bigram TF-IDF

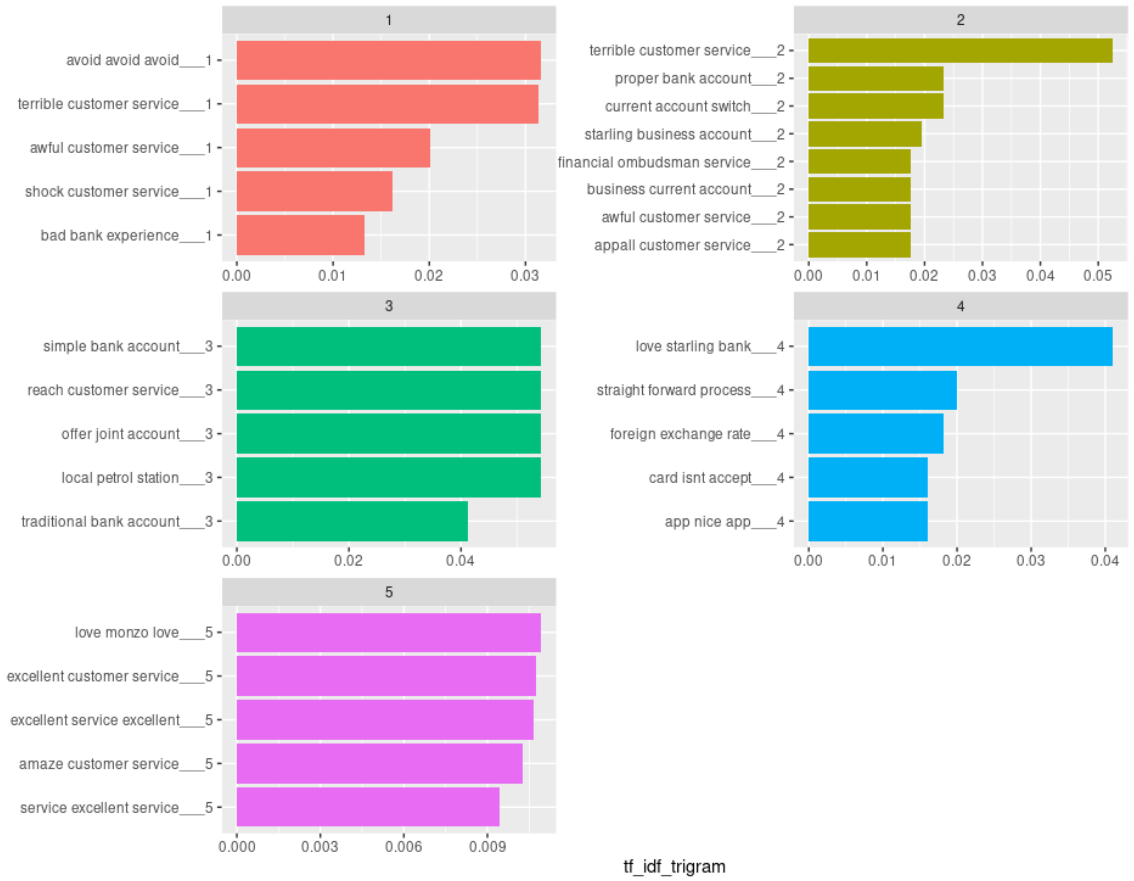


Figure A- 6 Digital Banks trigram TF-IDF

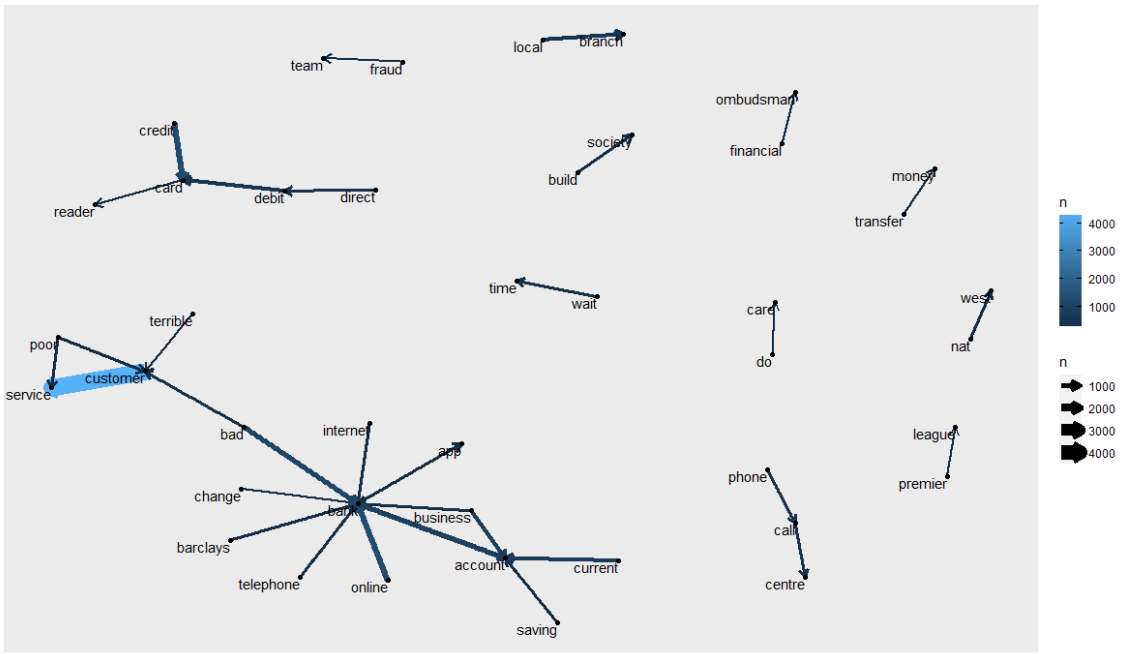


Figure A- 7 Traditional Banks word network

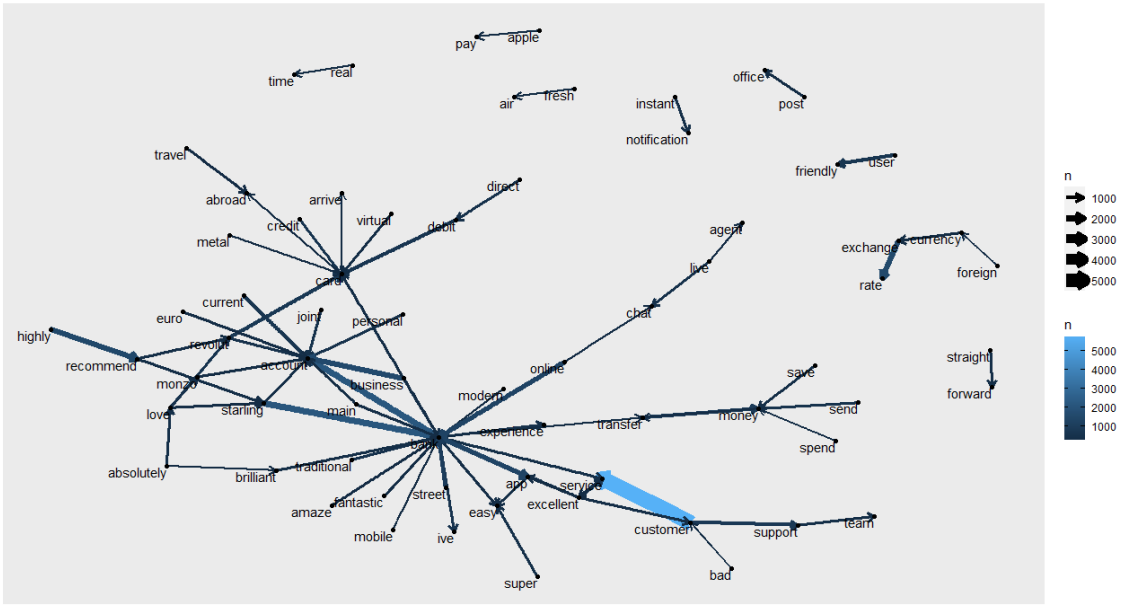


Figure A- 8 Digital Banks word network

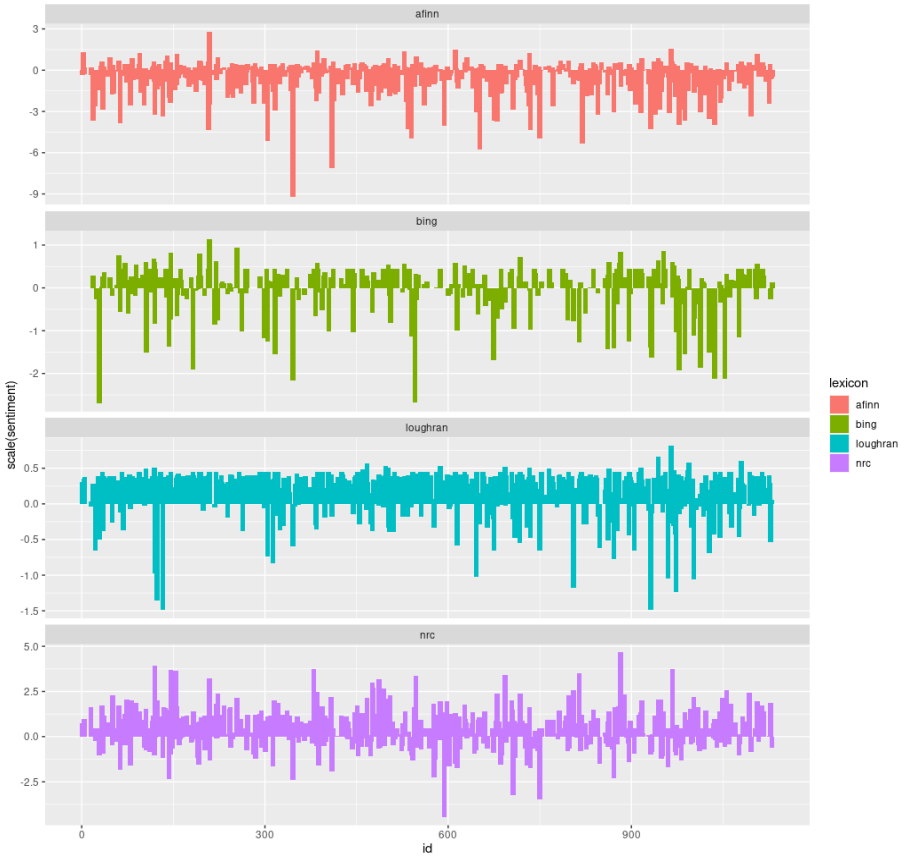


Figure A- 9 Traditional Banks 1 score polarity - tidytext

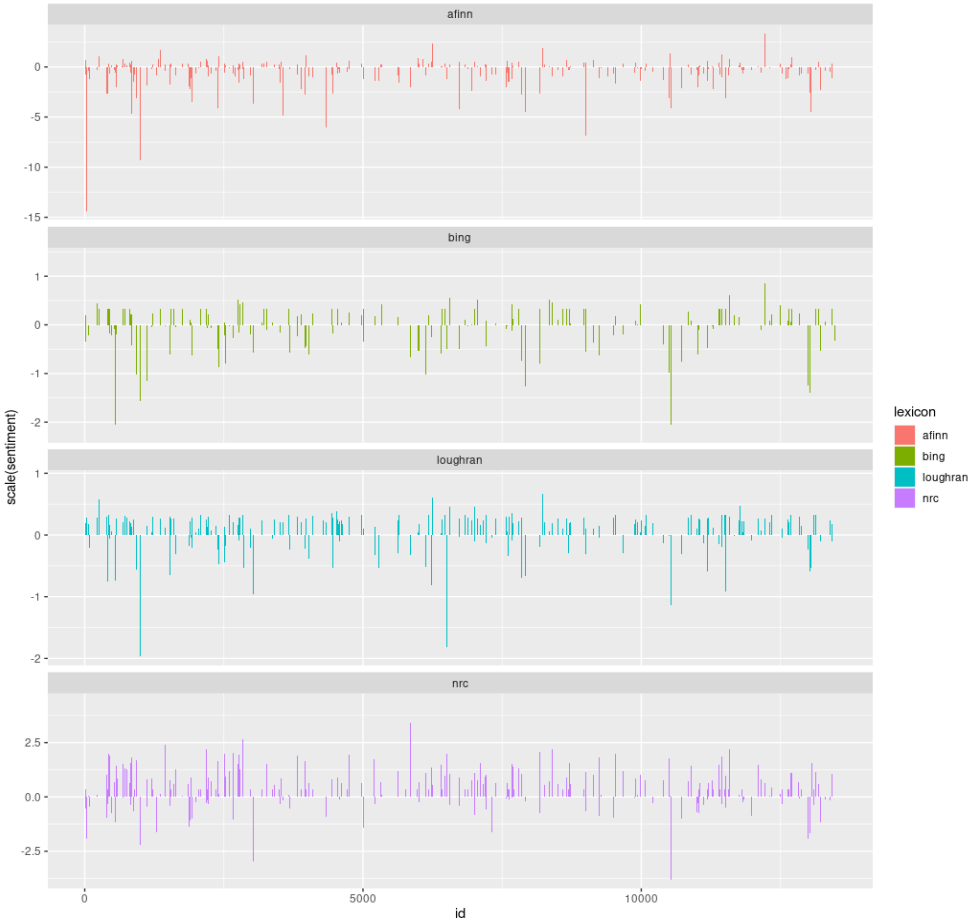


Figure A- 10 Traditional Banks 5 score polarity - tidytext

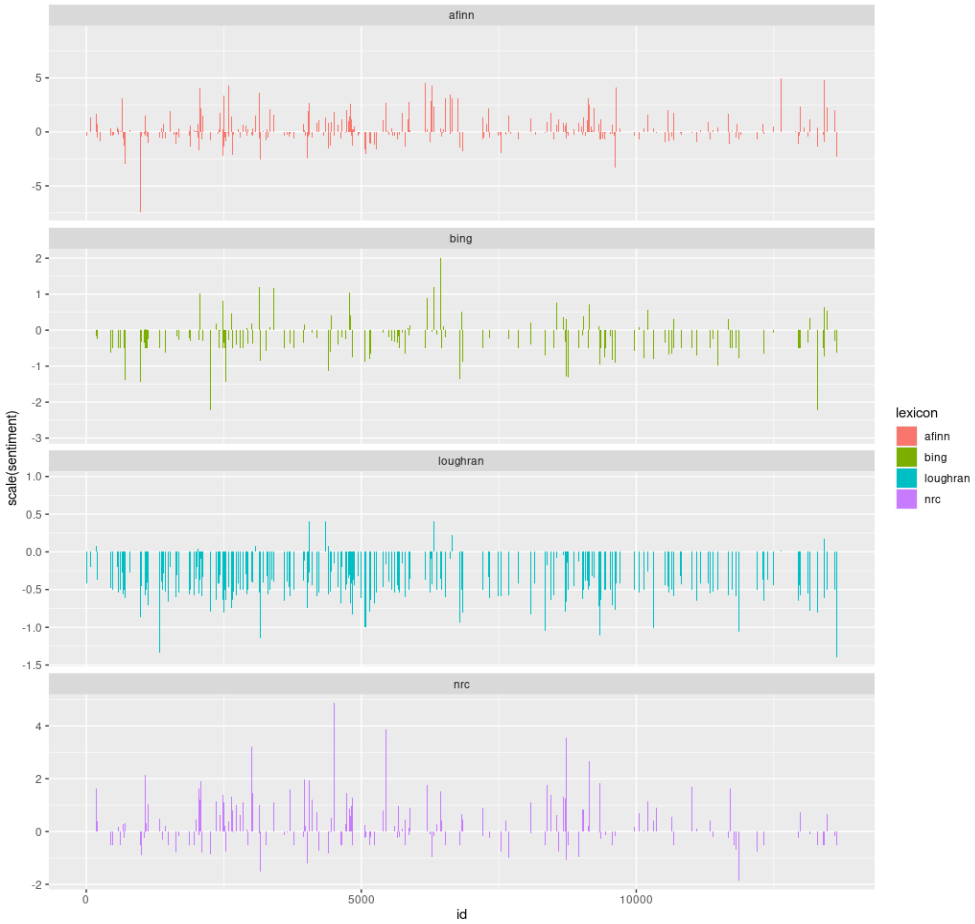


Figure A- 11 Digital Banks 1 score polarity - tidytext

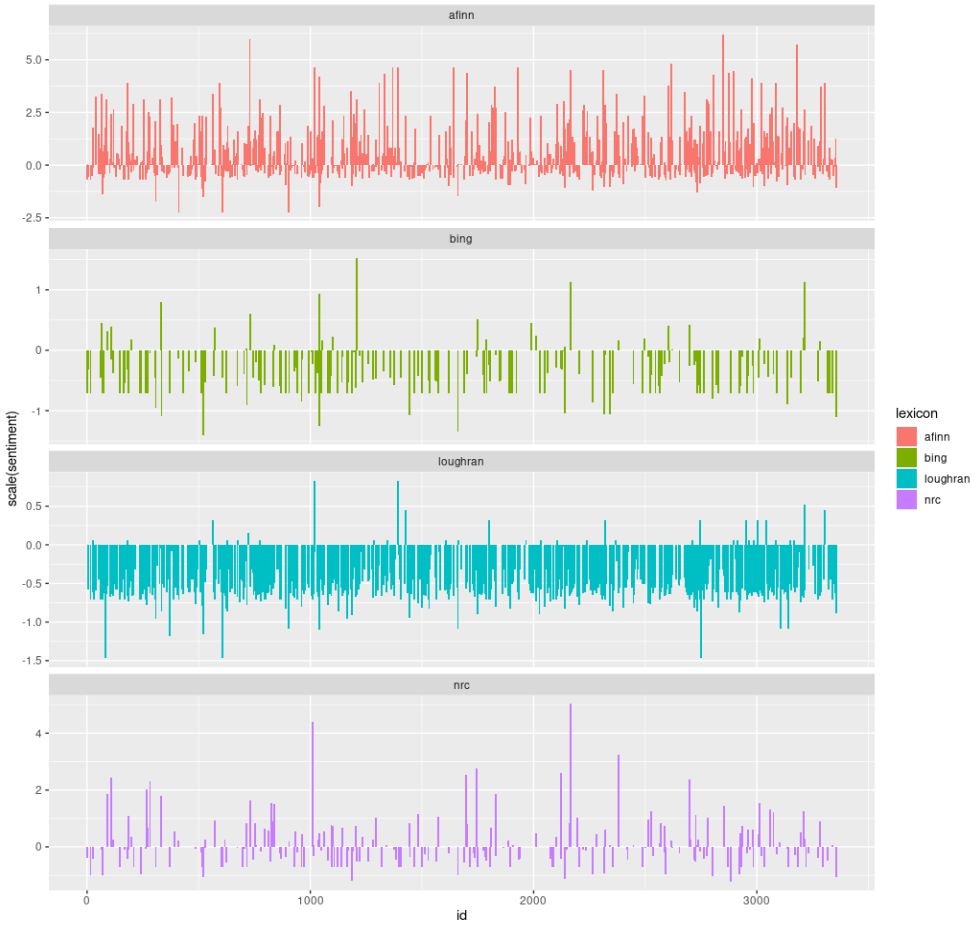


Figure A- 12 Digital Banks 5 score polarity – tidytext

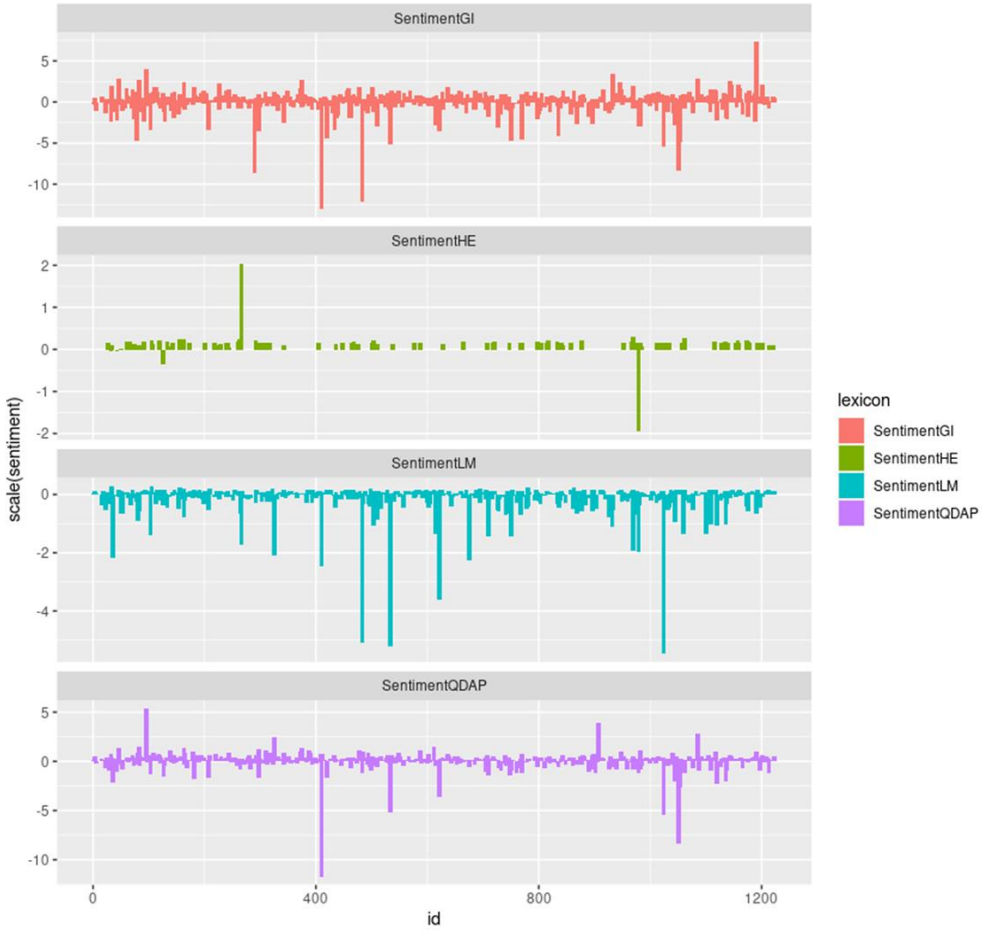


Figure A- 13 Traditional Banks 1 score polarity - SentimentAnalysis

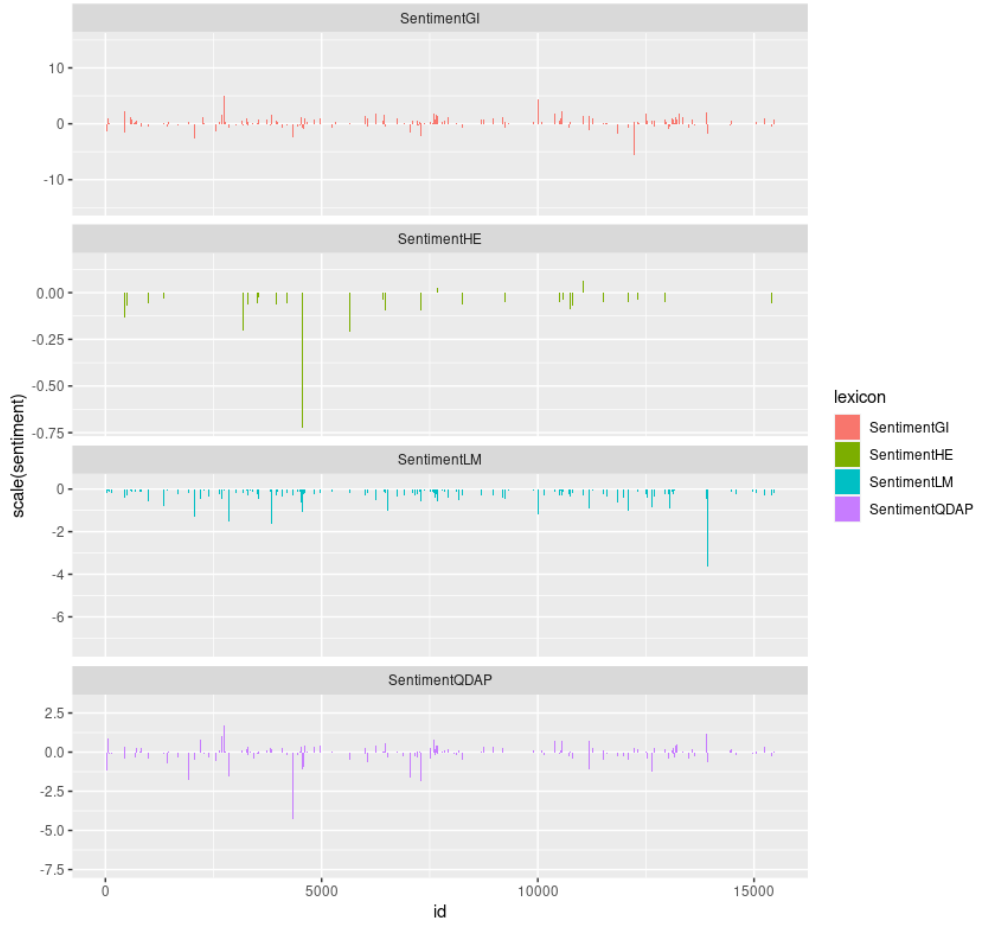


Figure A- 14 Traditional Banks 5 score polarity - SentimentAnalysis

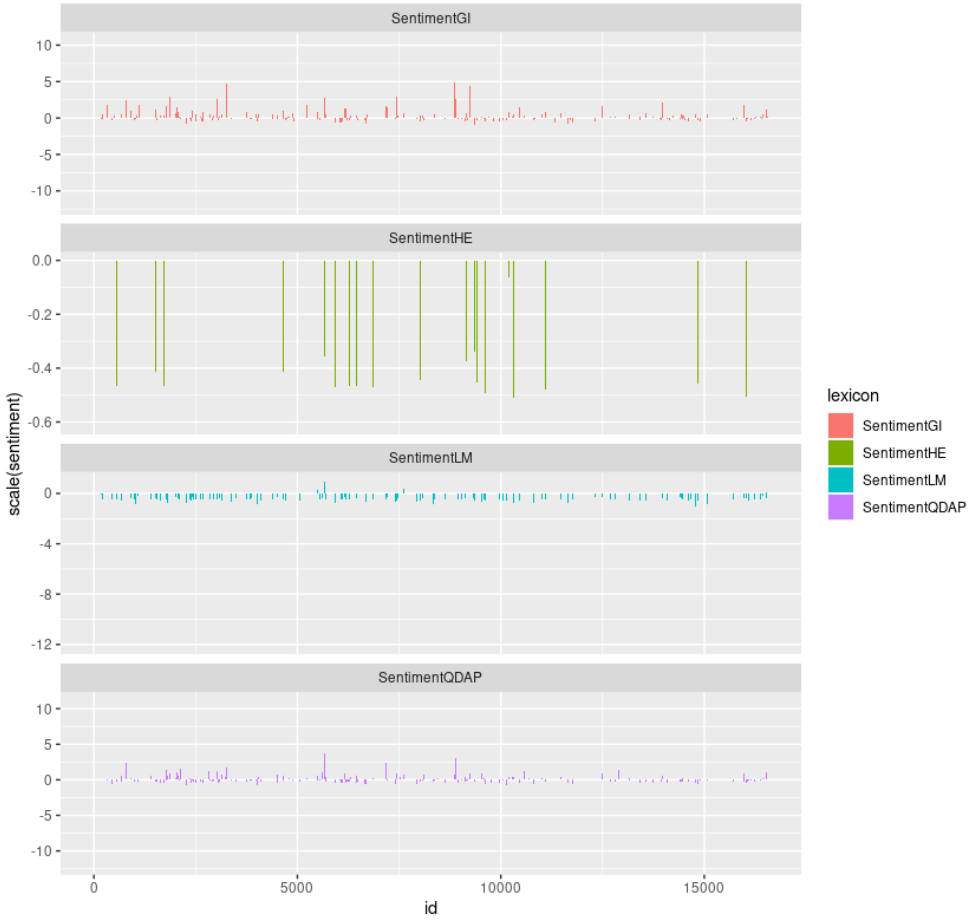


Figure A- 15 Digital Banks 1 score polarity - SentimentAnalysis

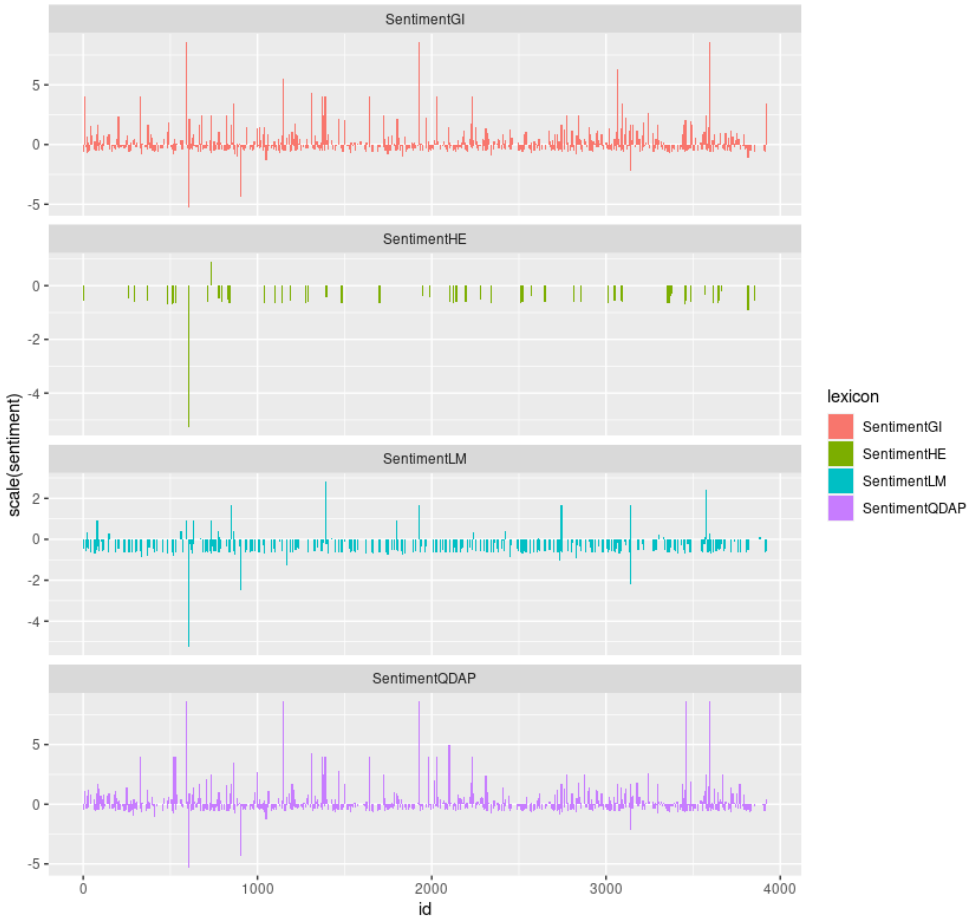


Figure A- 16 Digital Banks 5 score polarity – SentimentAnalysis

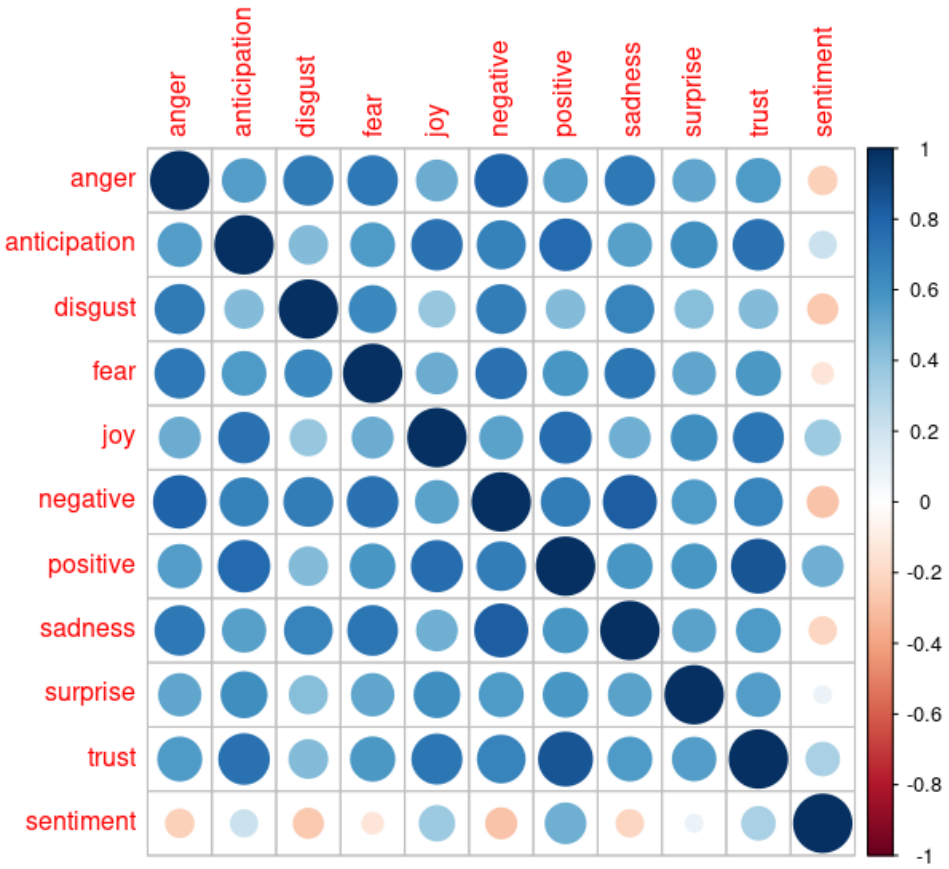


Figure A- 17 Traditional Bank NRC correlation

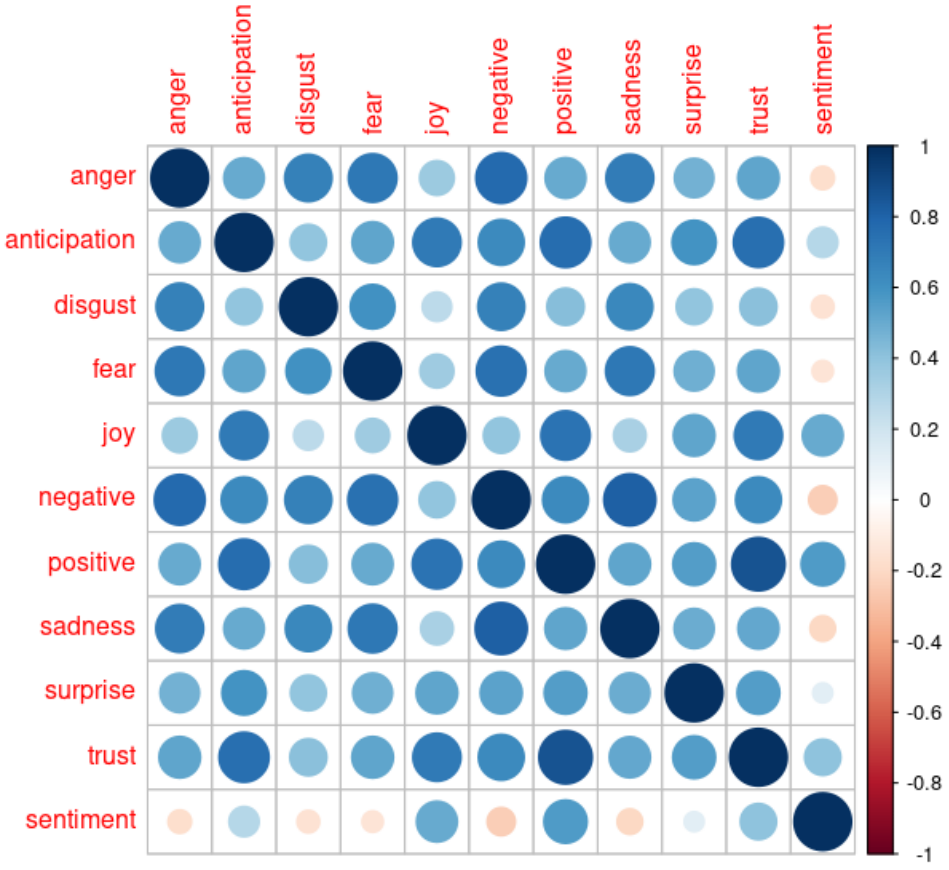


Figure A- 18 Digital Bank NRC correlation

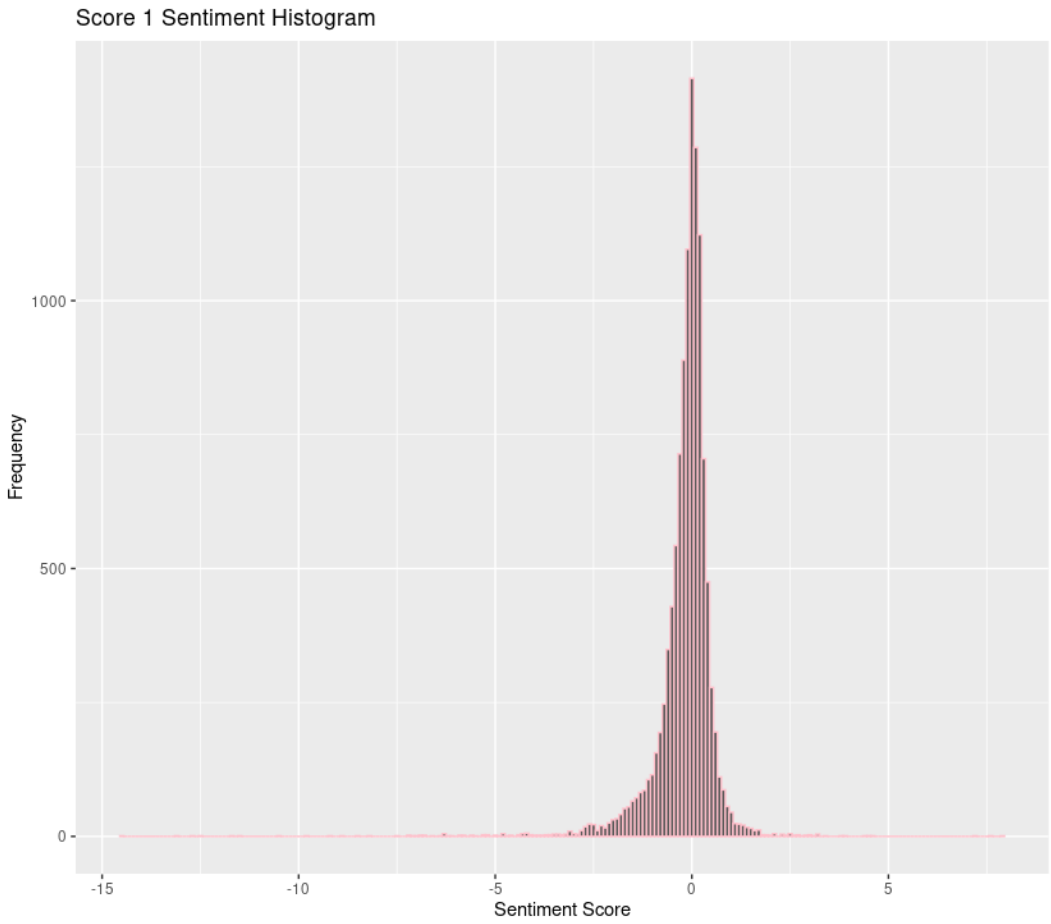


Figure A- 19 Traditional Banks 1 score - sentimentr

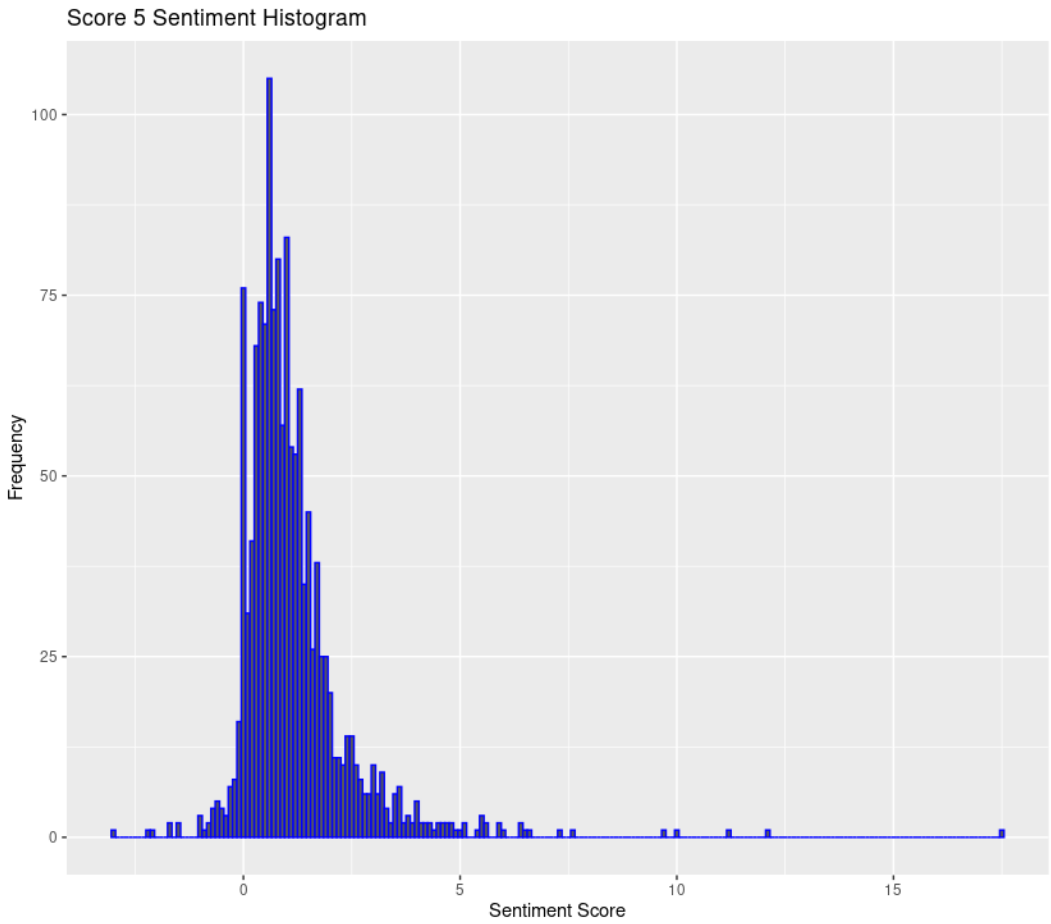


Figure A- 20 Traditional Banks 5 score - sentime

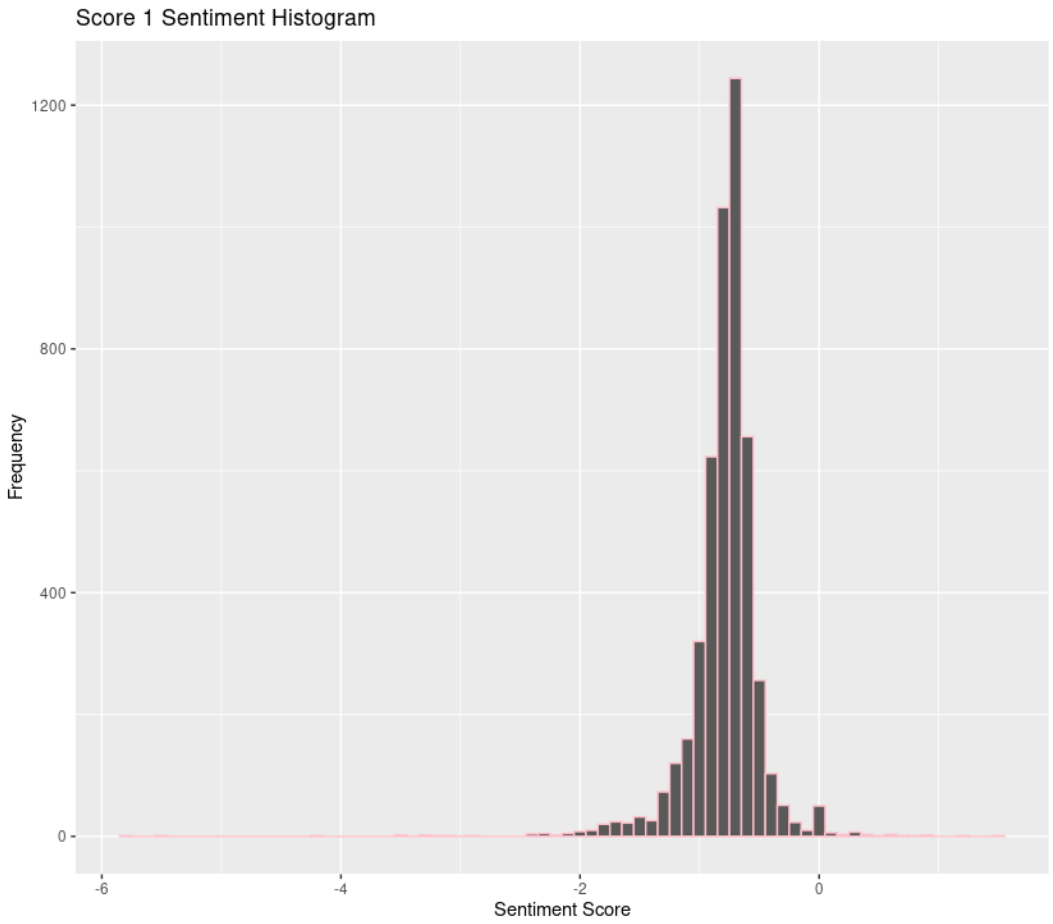


Figure A- 21 Digital Banks 1 score - sentimentr

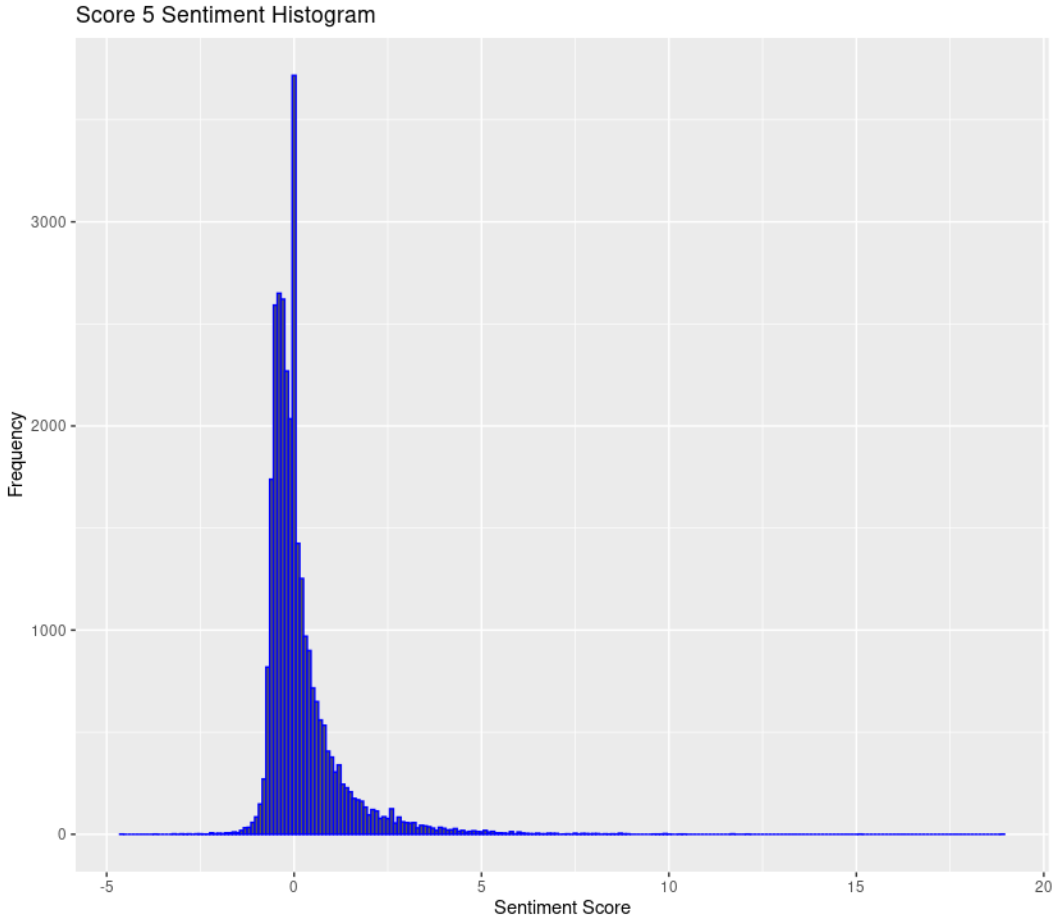


Figure A- 22 Digital Banks 5 score - sentimentr

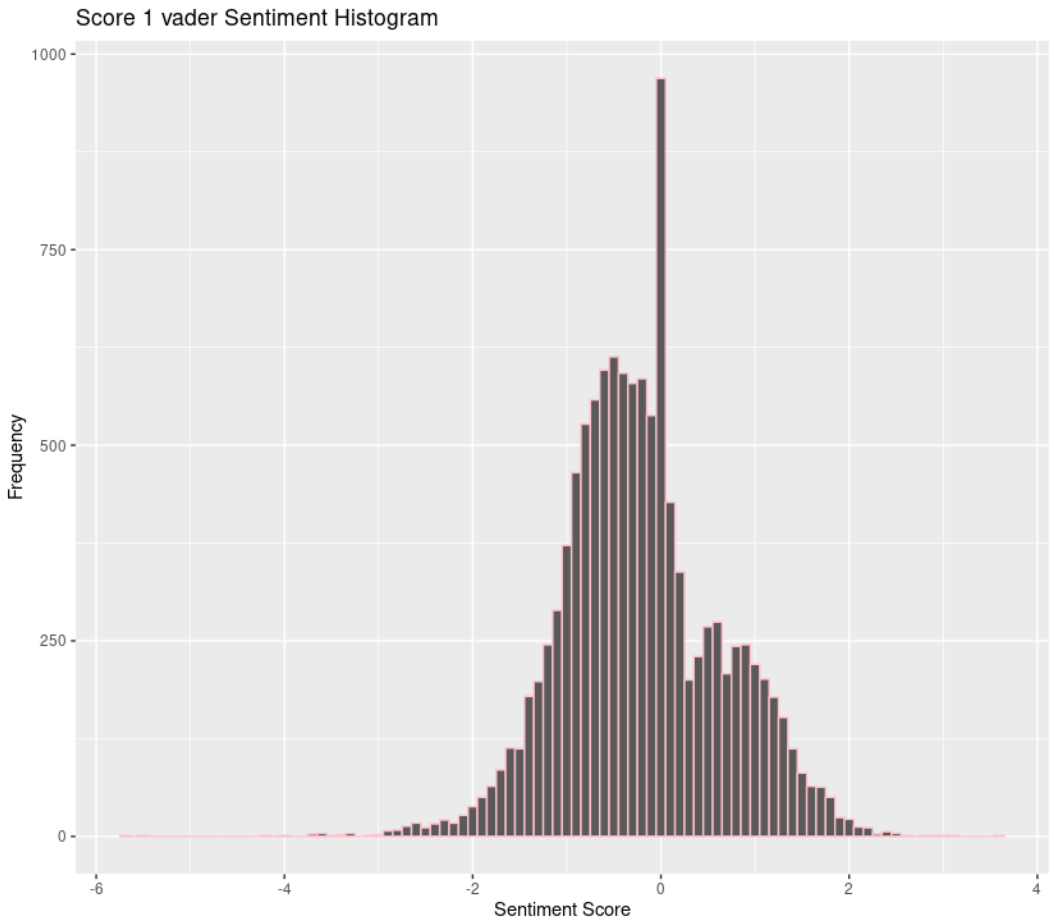


Figure A- 23 Traditional Banks 1 score - vader

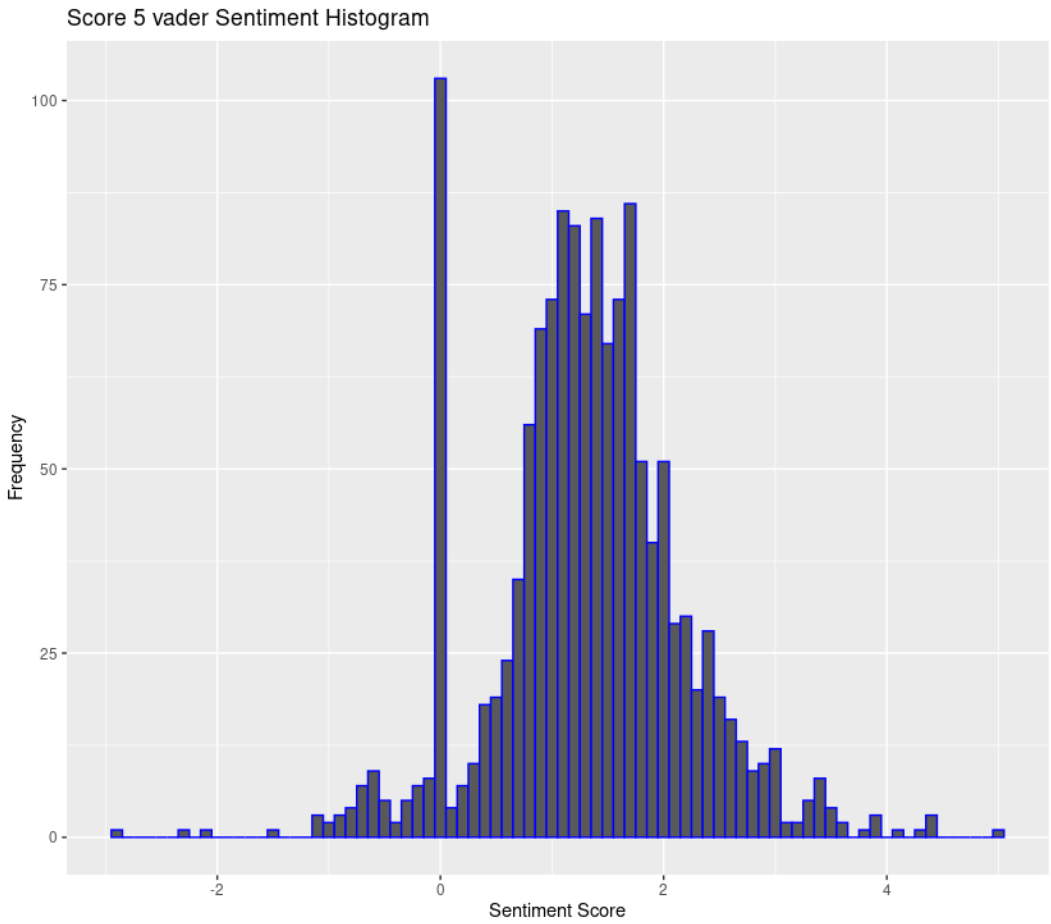


Figure A- 24 Traditional Banks 5 score - vader

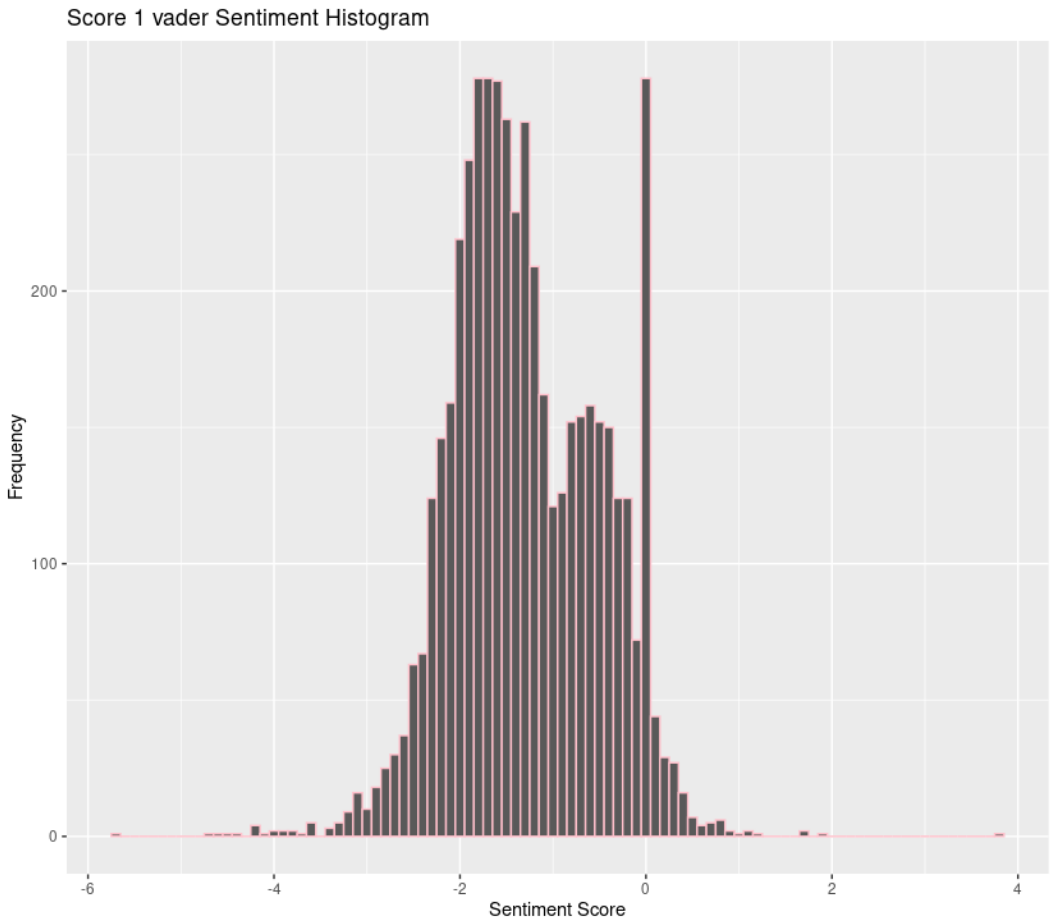


Figure A- 25 Digital Banks 1 score - vader

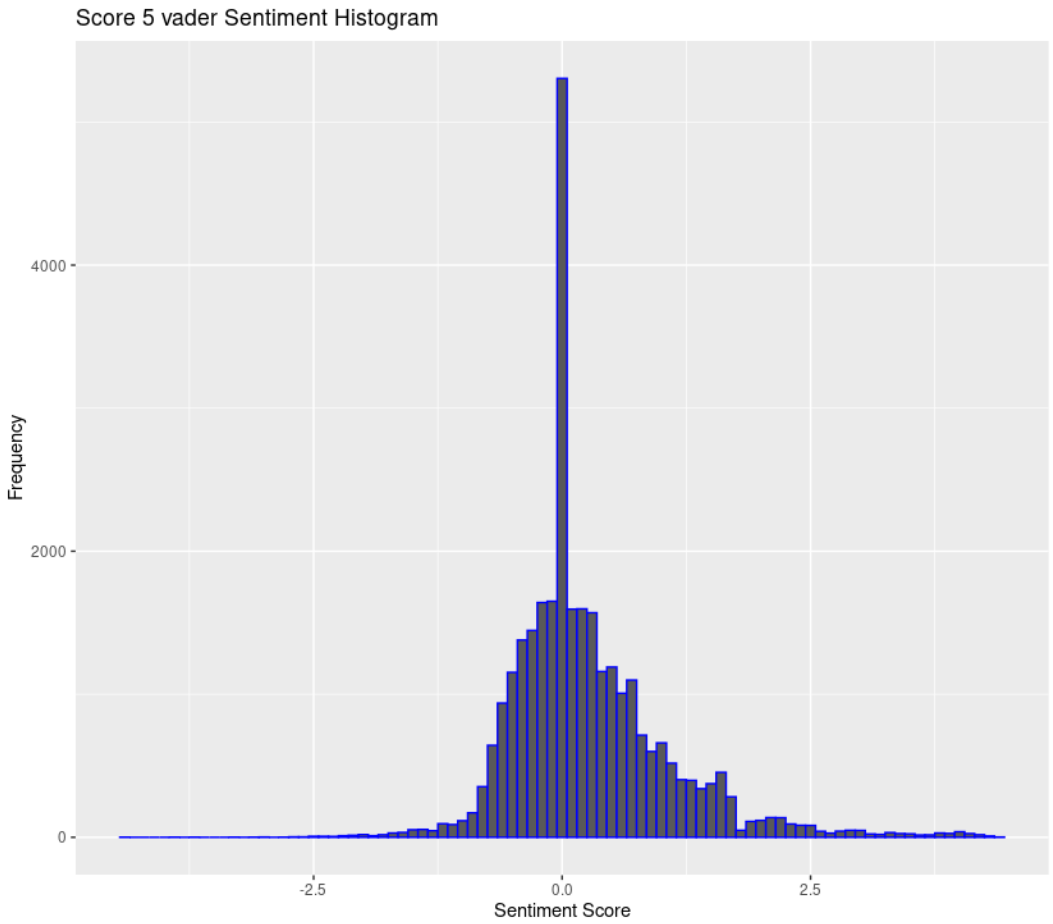
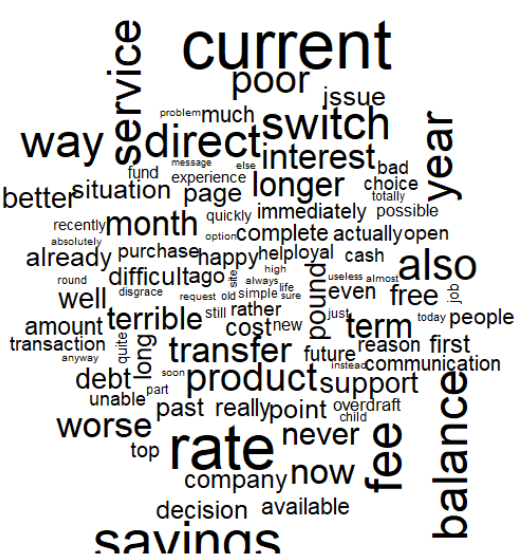
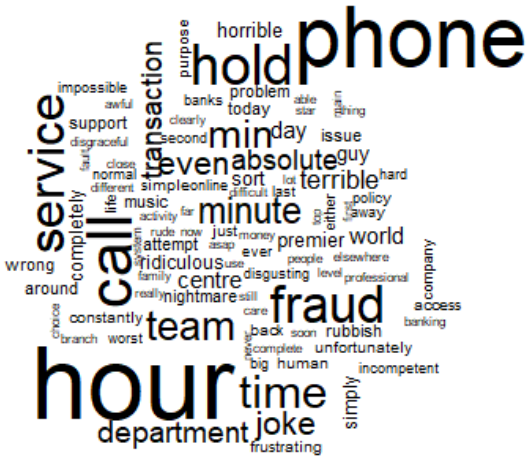
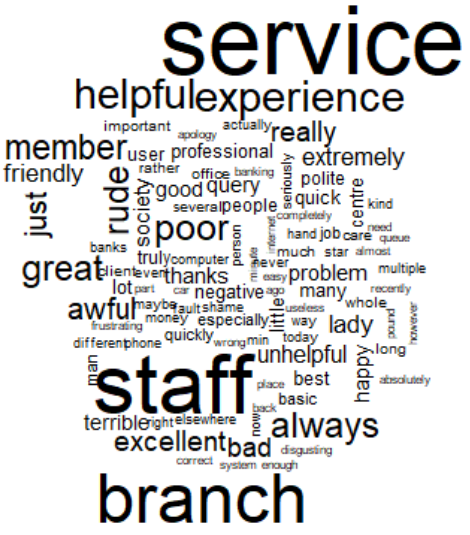


Figure A- 26 Digital Banks 5 score - vader







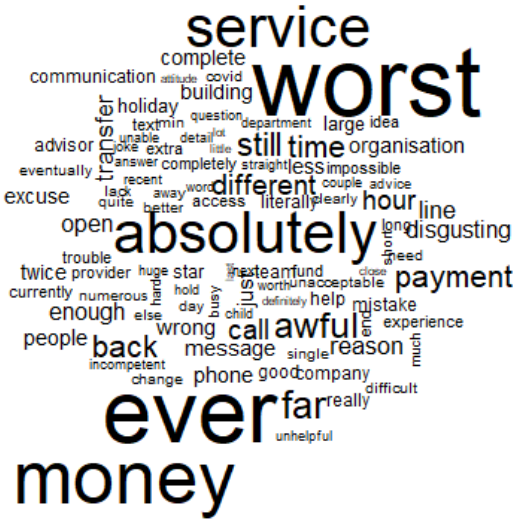
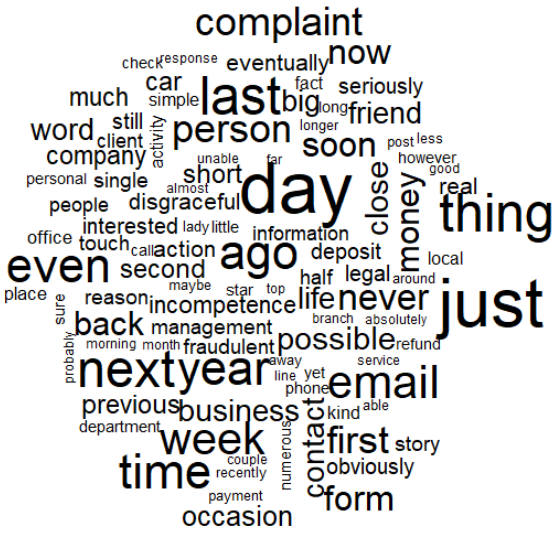
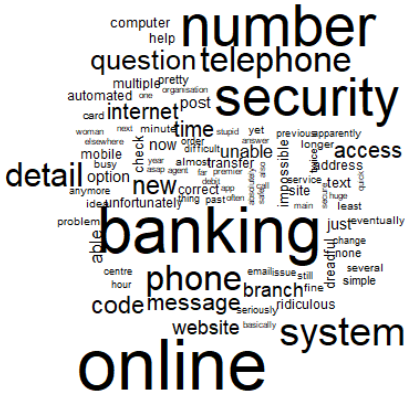
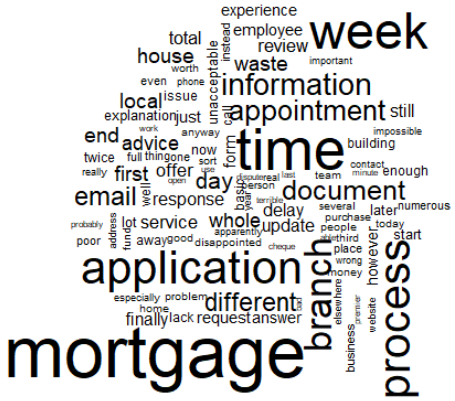
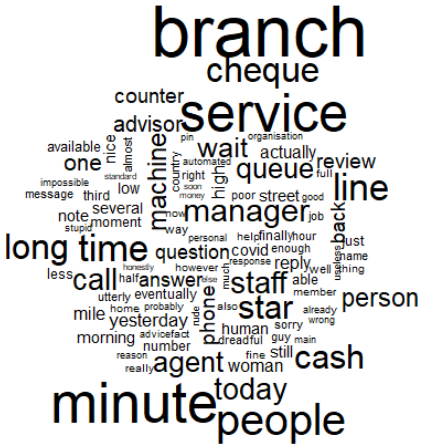


Figure A- 27 Traditional Banks word cloud topic 1-6









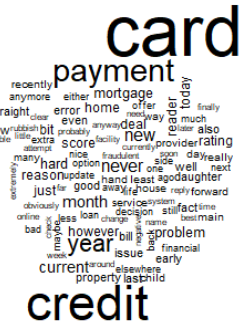


Figure A- 28 Traditional Banks Word Cloud topic 7 - 14

Figure A- 29 Digital Banks Word Cloud topic 1-7

Figure A- 30 Digital banks word cloud topic 8 – 14

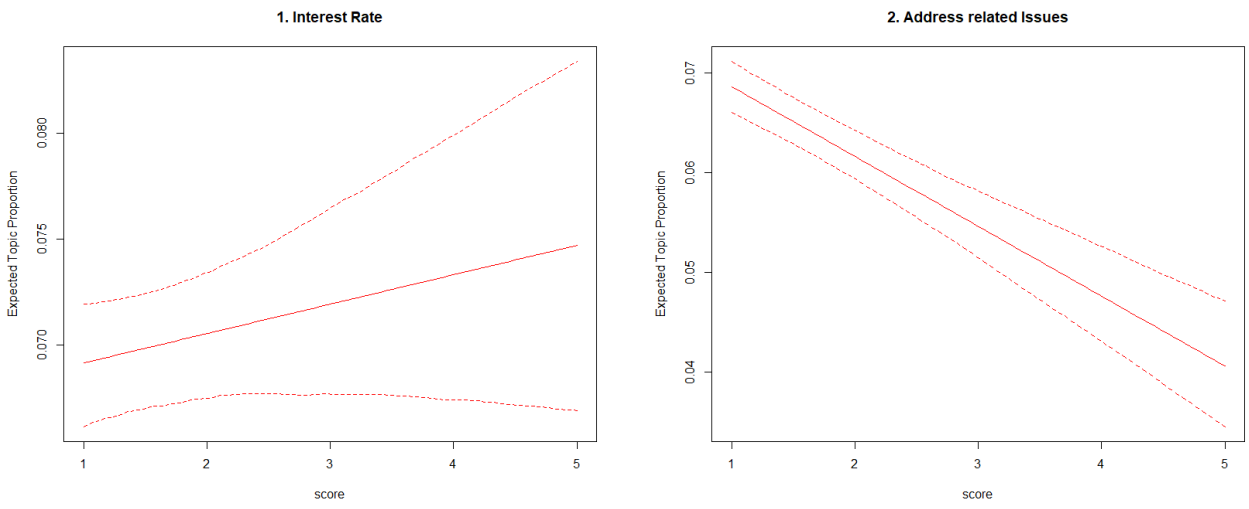


Figure A- 31 Traditional Banks Individual Marginal Effects topic 1-2

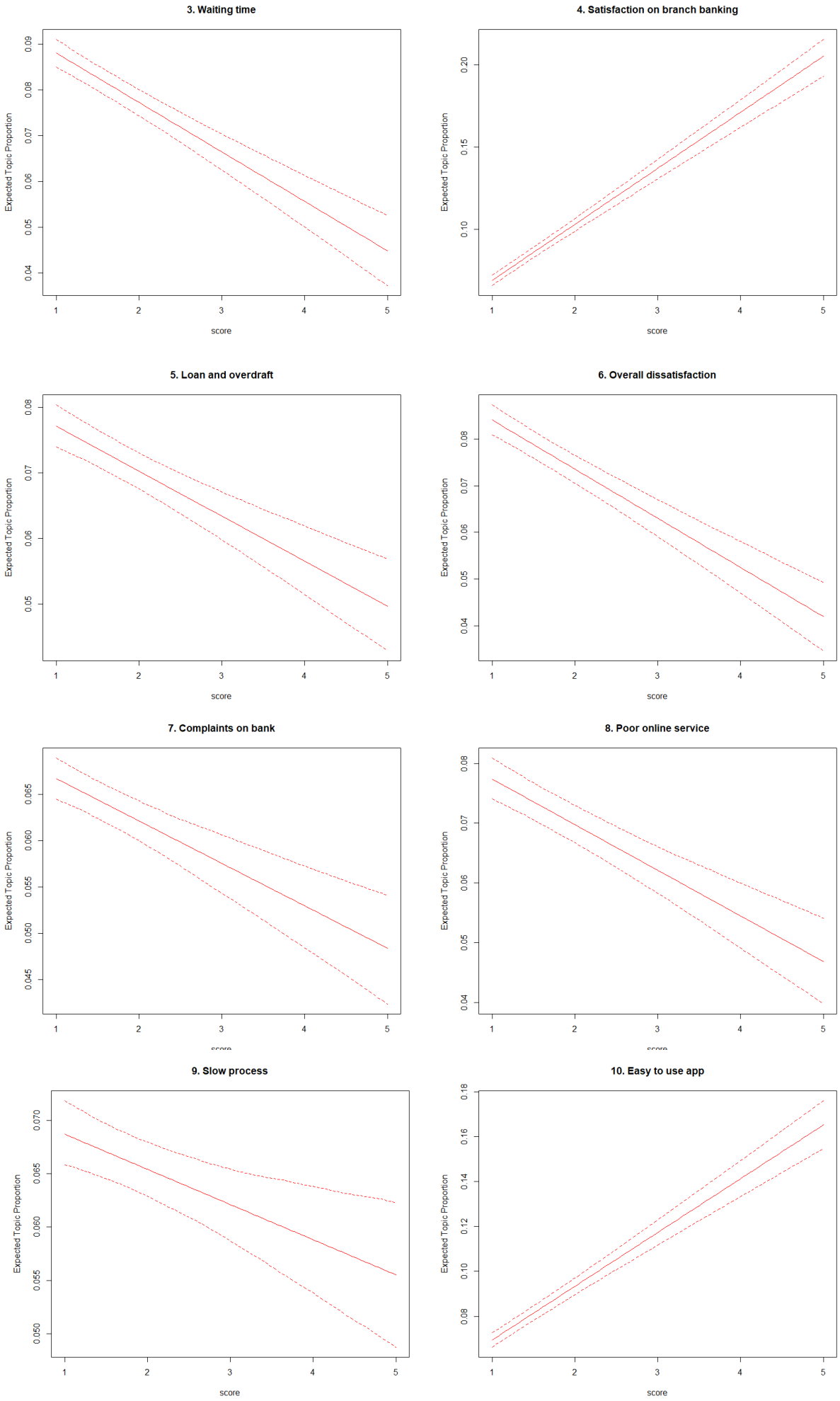


Figure A- 32 Traditional Banks Individual Marginal Effects topic 3-10

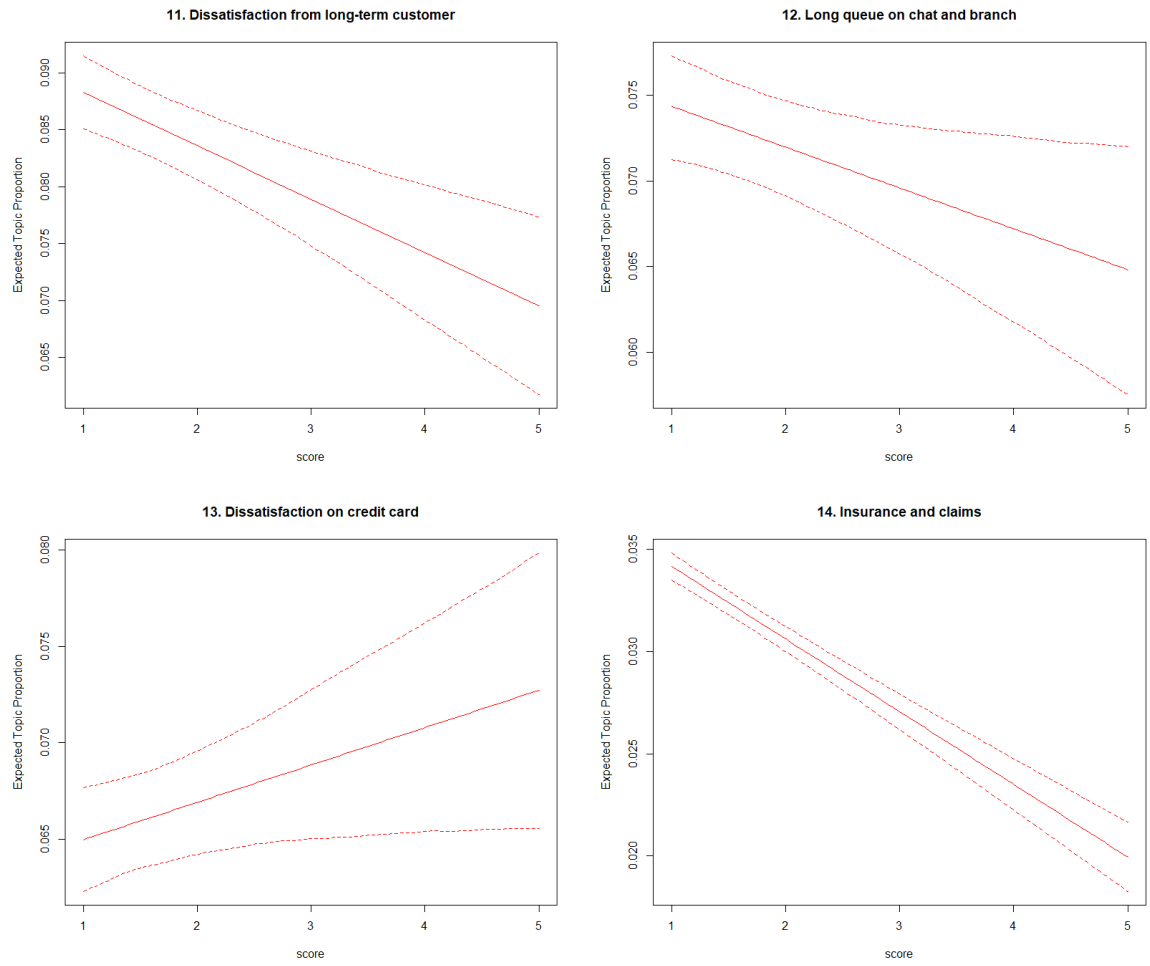


Figure A- 33 Traditional Banks Individual Marginal Effects topic 11-14

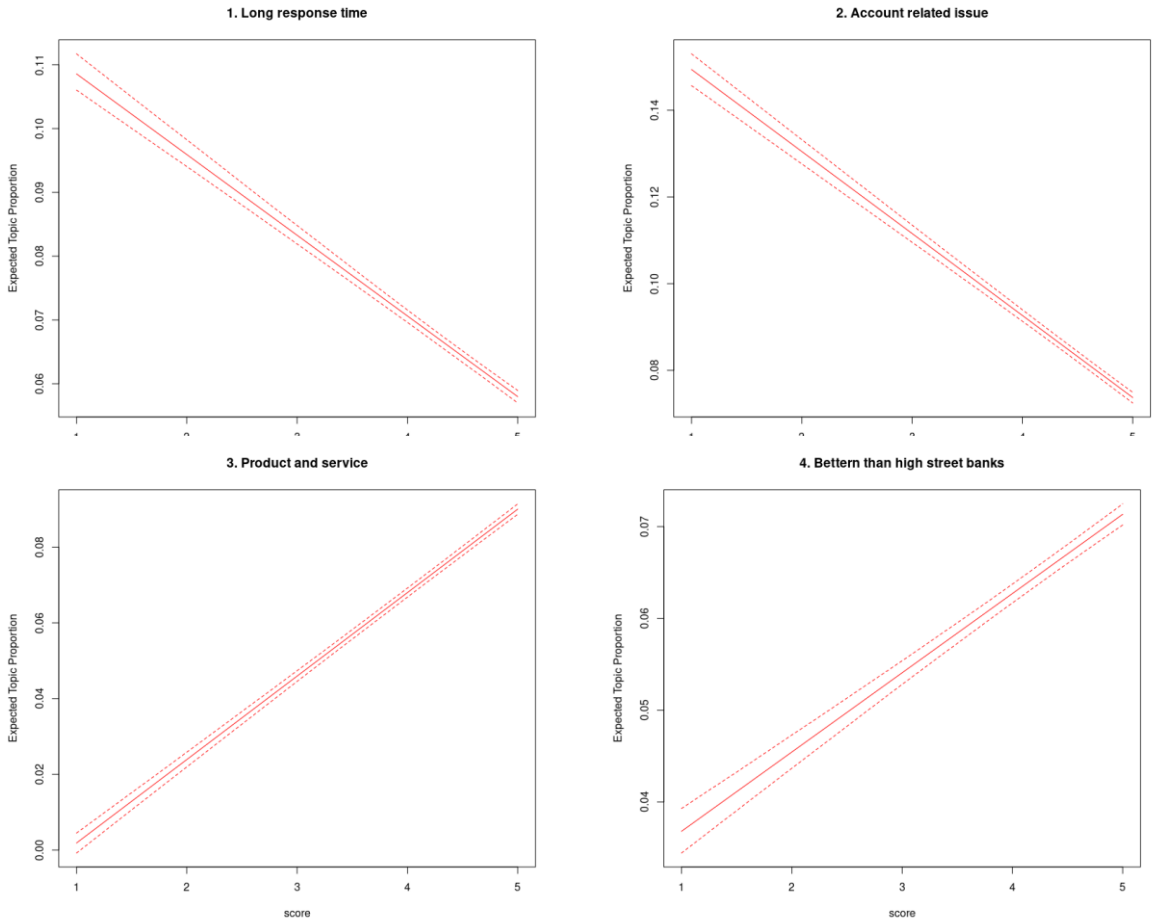
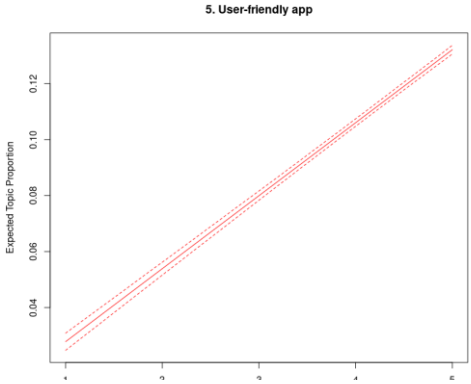
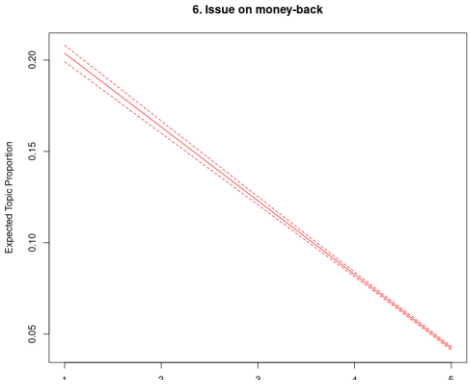
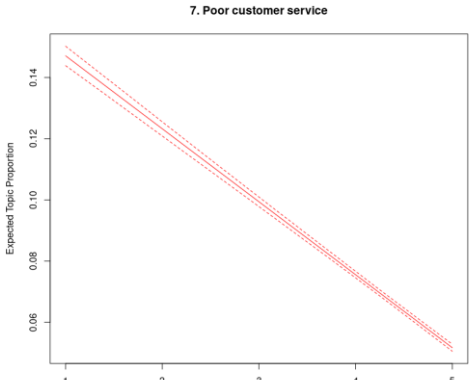
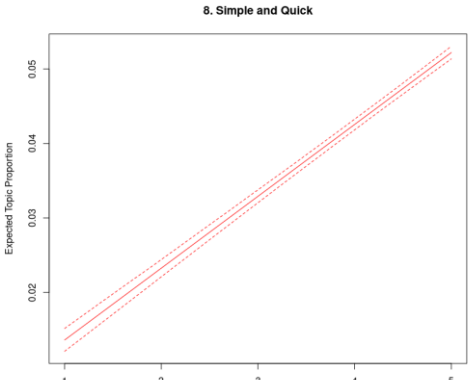


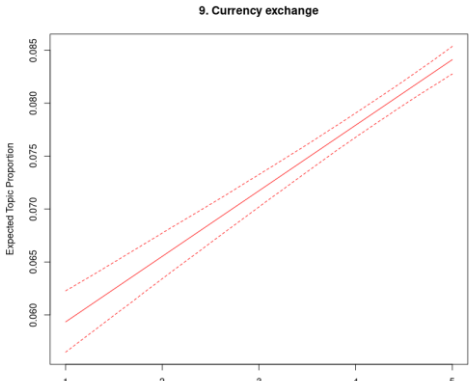
Figure A- 34 Digital Banks Individual Marginal Effects topic 1-4

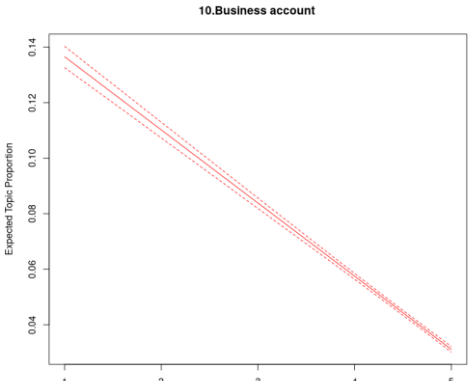

5. User-friendly app
Expected Topic Proportion
0.04
0.06
0.08
0.10
0.12

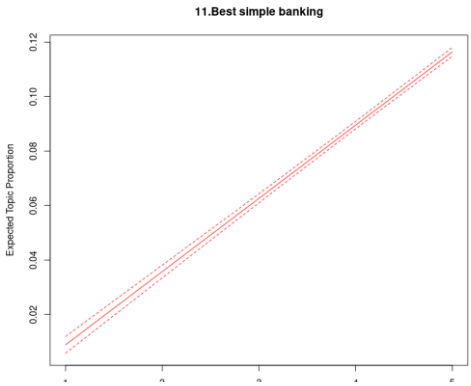

6. Issue on money-back
Expected Topic Proportion
0.05
0.10
0.15
0.20

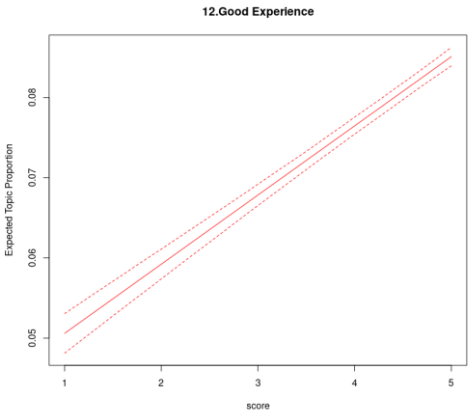

7. Poor customer service
Expected Topic Proportion
0.06
0.08
0.10
0.12
0.14

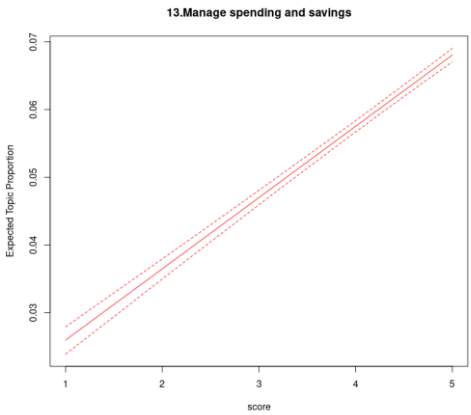

8. Simple and Quick
Expected Topic Proportion
0.02
0.03
0.04
0.05


9. Currency exchange
Expected Topic Proportion
0.060
0.065
0.070
0.075
0.080
0.085


10.Business account
Expected Topic Proportion
0.04
0.06
0.08
0.10
0.12
0.14


11.Best simple banking
Expected Topic Proportion
0.02
0.04
0.06
0.08
0.10
0.12


12.Good Experience
Expected Topic Proportion
0.05
0.06
0.07
0.08
1
2
3
4
5
score


13.Manage spending and savings
Expected Topic Proportion
0.03
0.04
0.05
0.06
0.07
1
2
3
4
5
score

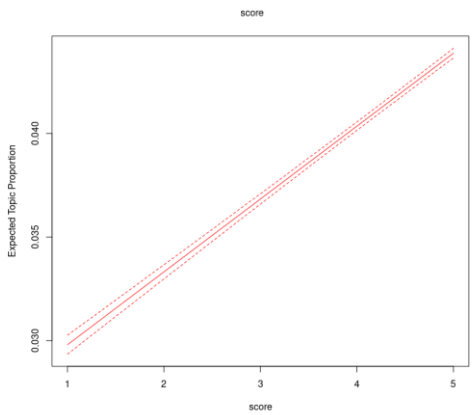

Expected Topic Proportion
0.030
0.035
0.040
1
2
3
4
5
score

Figure A- 35 Digital Banks Individual Marginal Effects topic 5-14

Traditional Banks

topic_label	proportion	frex_words
1. Interest Rate	6.65	fee, switch, savings, product, rate, balance, current
2. Address related Issues	6.3	debit, later, old, address, late, statement, course
3. Waiting time	8.44	fraud, hold, hour, min, joke, world, absolute
4. Satisfaction on branch banking	8.93	helpful, friendly, staff, excellent, rude, member, extremely
5. Loan and overdraft	7.22	charge, loan, overdraft, small, fund, monthly, money
6. Overall dissatisfaction	7.78	worst, ever, absolutely, far, awful, disgusting, open
7. Complaints on bank	6.58	next, friend, occasion, car, close, thing, big
8. Poor online service	6.58	security, number, telephone, system, code, detail, internet
9. Slow process	6.47	application, process, appointment, document, mortgage, waste, week
10. Easy to use app	7.78	app, easy, mobile, password, banks, good, issue
11. Dissatisfaction from long-term customer	8.73	personal, corrupt, totally, business, else, incompetent, level
12. Long queue on chat and branch	7.05	cheque, machine, cash, agent, wait, line, minute
13. Dissatisfaction on credit card	6.12	credit, card, rating, score, reader, payment, hard
14. Insurance and claims	5.39	claim, insurance, case, complaint, ombudsman, letter, matter

Digital Banks

topic_label	proportion	frex_words
1. Long response time	6.73	problem, support, payment, agent, team, system, first
2. Account related issue	8.94	year, now, current, account, main, month, yet
3. Product and service	7.06	excellent, amazing, brilliant, awesome, service, product, handy
4. Bettern than high street banks	6.09	high, spending, street, instant, notification, cheque, better
5. User-friendly app	12.11	user, great, friendly, app, straight, forward, way
6. Issue on money-back	6.95	back, people, fund, still, away, access, even
7. Poor customer service	6.47	phone, help, poor, new, customer, line, financial
8. Simple and Quick	4.26	quick, perfect, always, reliable, pleased, life, old
9. Currency exchange	7.95	exchange, rate, fee, currency, abroad, card, foreign
10.Business account	4.26	business, personal, email, information, response, process, company
11.Best simple banking	9.83	best, simple, banking, super, modern, efficient, fast
12.Good Experience	8.21	far, good, happy, experience, nice, impressed, helpful
13.Manage spending and savings	5.9	bill, easier, savings, fantastic, future, direct, space
14.Better than traditional bank	5.26	different, country, lot, also, mobile, bit, international

Figure A- 36 Frex words of 14 topics