

Trabajo Fin de Grado

Análisis de sentimiento en las cartas del CEO a sus accionistas

Álvaro Barrio Hernández

RESUMEN

La elaboración de este Trabajo Fin de Grado consiste en un análisis del texto presente en las carta del CEO a sus accionistas. La principal motivación es descubrir que existe más allá de las palabras, poder entender el sentimiento transmitido e inferir una serie de conclusiones acerca de los documentos. El análisis o minería de texto, considerado un subtipo de minería de datos, nos permite resaltar la información que a simple vista resulta difícil de encontrar.

Para la realización de este trabajo se han utilizado las 43 cartas enviadas anualmente (1977-2019) por Warren Buffett a sus accionistas. Utilizando técnicas basadas en estadística y con aplicación en el programa R-Studio se han podido tratar correctamente estos documentos para realizar el posterior análisis.

Primero comenzaremos con la obtención de los documentos y una primera visualización general, permitiéndonos conocer el contenido de los mismos. Continuaremos con un análisis más detallado mediante la distribución por años y la frecuencia de los tokens. A continuación nos adentraremos en el análisis de sentimiento, aplicando los diferentes léxicos para comparar entre los diferentes años y entre las diferentes unidades de palabras. Más adelante realizaremos un análisis por conjuntos de dos y tres palabras. Para finalizar aplicaremos técnicas de clasificación inversa y agrupación no supervisada permitiéndonos llegar a resultados no triviales a simple vista.

Nos enfocaremos en contestar algunas preguntas como las siguientes:

- ¿Cómo se presentan los datos a tratar?
- ¿Qué sentimiento presentan las cartas del señor Warren Buffett?
- ¿Existen variaciones significativas del contenido en función de los años?
- ¿Qué palabras contribuyen a generar los diferentes sentimientos?
- ¿Existen relaciones entre las propias palabras?
- ¿Cómo podemos conocer el tema principal de nuestras cartas?

ÍNDICE

Capítulos

1.	<i>INTRODUCCIÓN</i>	8
2.	<i>OBTENCIÓN DE DOCUMENTOS Y PRIMER VISUALIZADO</i>	9
3.	<i>MÁS ALLÁ DE LAS PALABRAS</i>	13
3.1	¿Qué años contienen las palabras más repetidas?.....	14
3.2	Frecuencias totales por cada año.....	16
3.3	Palabras más repetidas en cada año.....	18
4.	<i>ANÁLISIS DE SENTIMIENTO A TRAVÉS DE LOS AÑOS</i>	20
4.1	Afinn.....	21
4.2	Loughran.....	24
4.3	Nrc.....	28
4.4	Bing	31
4.5	Combinación de los 4 léxicos	35
5.	<i>RELACIONES ENTRE PALABRAS</i>	36
5.1	Utilizar Bigrams para obtener sentimientos en contexto	37
5.2	Aplicando “graph” a nuestros bigrams.....	39
5.3	Correlación de pares no consecutivos	41
5.3.1	Recuento y correlación entre secciones.....	42
5.3.2	Correlación Pairwise	43
5.4	Trigrams.....	45
6.	<i>INVERSE DOCUMENT FREQUENCY</i>	46
6.1	Frecuencia de término en las cartas del señor Warren buffet.....	46
6.2	Ley de Zipf	48
6.3	La función bind_tf_idf.....	50
7.1	LDA en años	55
7.2	Probabilidades de pertenencia a un tema.....	60
7.3	Asignaciones de palabras	62
8.	<i>CONCLUSIONES</i>	64
9.	<i>BIBLIOGRAFÍA</i>	65

Índice de Figuras

FIGURA 1: CADENA DE TRANSFORMACIONES	9
FIGURA 2: PALABRAS MÁS FRECUENTES EN %	10
FIGURA 3: WORDCLOUD DE 100 PALABRAS	11
FIGURA 4: WORDCLOUD DE 200 PALABRAS	12
FIGURA 5: PALABRAS MÁS REPETIDAS DISTRIBUIDAS EN AÑOS (N MAYOR A 50)	14
FIGURA 6: PALABRAS MÁS REPETIDAS DISTRIBUIDAS EN AÑOS (N MAYOR A 60)	15
FIGURA 7: PALABRAS MÁS REPETIDAS DISTRIBUIDAS EN AÑOS (N MAYOR A 70)	15
FIGURA 8: PALABRAS TOTALES POR AÑOS. GRÁFICO DE LÍNEAS	17
FIGURA 9: PALABRAS TOTALES POR AÑOS. GRÁFICO DE TENDENCIA	17
FIGURA 10: TOP 5 PALABRAS MÁS REPETIDAS EN CADA AÑO	19
FIGURA 11: DISTRIBUCIÓN DE PALABRAS POR SENTIMIENTO - AFINN (TOP 30)	22
FIGURA 12: DISTRIBUCIÓN POR AÑOS - SENTIMENTO AFINN (BARRAS)	23
FIGURA 13: DISTRIBUCIÓN POR AÑOS - SENTIMENTO AFINN (PUNTOS Y TENDENCIA)	24
FIGURA 14: CLASIFICACIÓN DE PALABRAS POR SENTIMIENTO - LOUGHAN	25
FIGURA 15: DISTRIBUCIÓN POR AÑOS - SENTIMENTO LOUGHAN	26
FIGURA 16: DISTRIBUCIÓN POR AÑOS - SENTIMENTO LOUGHAN (DIVIDIDO POR SENTIMIENTO)	27
FIGURA 17: DISTRIBUCIÓN POR AÑOS - SENTIMENTO LOUGHAN (DIVIDIDO POR AÑOS)	27
FIGURA 18: CLASIFICACIÓN DE PALABRAS POR SENTIMIENTO - NRC	29
FIGURA 19: DISTRIBUCIÓN POR AÑOS - SENTIMENTO NRC	30
FIGURA 20: DISTRIBUCIÓN POR AÑOS - SENTIMENTO NRC (DIVIDIDO POR SENTIMIENTO)	30
FIGURA 21: DISTRIBUCIÓN POR AÑOS - SENTIMENTO NRC	31
FIGURA 22: NUBE DE PALABRAS - SENTIMENTO NRC	33
FIGURA 23: DISTRIBUCIÓN POR AÑOS - SENTIMENTO NRC	33
FIGURA 24: DISTRIBUCIÓN POR AÑOS - SENTIMENTO NRC (DIVIDIDO POR SENTIMIENTO)	34
FIGURA 25: DISTRIBUCIÓN POR AÑOS - SENTIMENTO NRC (DIVIDIDO POR AÑO)	34
FIGURA 26: APLICACIÓN DE LOS 4 LÉXICOS	35
FIGURA 27: TOP 20 PALABRAS SEGUIDAS DE 'NOT'	38
FIGURA 28: PALABRAS QUE MÁS CONTRIBUYEN AL SENTIMIENTO NEGATIVO	38
FIGURA 29: APLICACIÓN DE GGRAPH A NUESTROS BIGRAMS.	39
FIGURA 30: APLICACIÓN DE GGRAPH MEJORADO A NUESTROS BIGRAMS	40
FIGURA 31: APLICACIÓN DE GGRAPH MEJORADO A NUESTROS BIGRAMS (2)	41
FIGURA 32: LA FILOSOFÍA DETRÁS DEL PAQUETE WIDYR	42
FIGURA 33: COEFICIENTE DE PHI	43
FIGURA 34: MAYORES CORRELACIONES DE 'CUSTOMER' Y 'FEAR'	44
FIGURA 35: MAYORES CORRELACIONES DE 'CUSTOMER' Y 'FEAR'	45
FIGURA 36: EXPRESIÓN MATEMÁTICA 'TF-IDF'	46
FIGURA 37: FRECUENCIA DE TÉRMINOS EN LAS CARTAS	47
FIGURA 38: FRECUENCIA DE TÉRMINOS EN CADA CARTA.	47
FIGURA 39: VISUALIZACIÓN LEY DE ZIPF	49
FIGURA 40: VISUALIZACIÓN LEY DE ZIPF POR CARTA	49
FIGURA 41: VISUALIZACIÓN LEY DE ZIPF AJUSTE CON UN EXPONENTE	50
FIGURA 42: VISUALIZACIÓN DE TF_IDF POR AÑO	51
FIGURA 43: VISUALIZACIÓN DE TF_IDF POR AÑO (ZOOM)	52
FIGURA 44: VISUALIZACIÓN DE TF_IDF POR AÑO (SIN STOPWORDS)	53
FIGURA 45: VISUALIZACIÓN DE TF_IDF POR AÑO (SIN STOPWORDS Y ZOOM)	53
FIGURA 46: DIAGRAMA DE FLUJO DE UN ANÁLISIS DE TEXTO	54

FIGURA 47: CLASIFICACIÓN EN 16 TEMAS (PARTE 1)	57
FIGURA 48: CLASIFICACIÓN EN 16 TEMAS (PARTE 2)	58
FIGURA 49: CLASIFICACIÓN EN 2 TEMAS	59
FIGURA 50: LAS PROBABILIDADES GAMMA PARA CADA TEMA	61
FIGURA 51: MATRIZ DE CLASIFICACIÓN	63

Índice de Tablas

TABLA 1: PALABRAS MÁS FRECUENTES	10
TABLA 2: PALABRAS MÁS REPETIDAS CON SUS RESPECTIVOS AÑOS	13
TABLA 3: PALABRAS MÁS REPETIDAS EN 1977 Y 2019	13
TABLA 4: SUMATORIO DE PALABRAS POR CADA AÑO	16
TABLA 5: LA PALABRA MÁS REPETIDA EN CADA AÑO	18
TABLA 6: LÉXICOS	20
TABLA 7: PALABRAS CLASIFICADAS - AFINN	21
TABLA 8: PALABRAS POSITIVAS CLASIFICADAS EN CADA AÑO - AFINN	22
TABLA 9: PALABRAS REGISTRADAS EN LOUGHAN	24
TABLA 10: CONTADOR DE PALABRAS CLASIFICADAS EN LOUGHAN (GENERAL VS APLICADO)	25
TABLA 11: CLASIFICACIÓN EN NRC	28
TABLA 12: CONTADOR DE PALABRAS REGISTRADAS EN NRC (GENERAL VS APLICADO)	28
TABLA 13: CLASIFICACIÓN EN BING	32
TABLA 14: CONTADOR DE PALABRAS REGISTRADAS EN BING (GENERAL VS APLICADO)	32
TABLA 15: BIGRAMS	36
TABLA 16: BIGRAMS FILTRADOS	37
TABLA 17: BIGRAMS APLICANDO GRAPH	39
TABLA 18: CORRELACIÓN ENTRE PALABRAS	42
TABLA 19: CORRELACIÓN ENTRE PALABRAS	43
TABLA 20: CORRELACIÓN ENTRE PALABRAS	44
TABLA 21: TOP 10 TRIGRAMS	45
TABLA 22: FRECUENCIA DE TÉRMINOS	46
TABLA 23: APLICACIÓN DE ZIPF	48
TABLA 24: EXPONENTE Y PENDIENTE	50
TABLA 25: TF_IDF	51
TABLA 26: CONTADOR POR AÑO	55
TABLA 27: CREACIÓN DE LA MATRIZ	55
TABLA 28: LDA	56
TABLA 29: BETAS DEL MODELO	56
TABLA 30: BETAS DEL MODELO	60
TABLA 31: CONSENSO	61
TABLA 32: FUNCIÓN AUGMENT()	62

1. INTRODUCCIÓN

Cada año, el señor Warren Buffett envía una carta pública a los accionistas de Berkshire Hathaway (conglomerado de empresas). En los documentos podemos encontrar información sobre el resultado de sus inversiones, así como previsiones sobre el propio mercado.

La principal característica de estos documentos es el modo en el que están escritos. La mayoría de reportes financieros suelen estar presentados de una forma densa, técnica e incluso en ocasiones, confusa. Nos encontramos con unas cartas fácilmente entendibles y accesibles para todos los públicos.

En este trabajo se realizará un análisis del sentimiento presente en las cartas de Warren Buffett a los accionistas entre 1977 y 2019. El análisis de sentimiento implica las técnicas necesarias para cuantificar el tipo de sentimiento general que está presente en un texto. El primer caso de uso que se nos viene a la mente puede ser determinar qué tan positivo o negativo es un documento concreto, pero comprobaremos que hay muchas más clasificaciones disponibles.

El uso de los datos ordenados es una forma poderosa de hacer que su manejo sea más fácil y más efectivo. La forma de presentación será posicionar cada variable en una columna y cada observación en una fila. En ocasiones se mencionara el término ‘token’ haciendo referencia a una unidad de texto significativa. La ‘tokenización’ es el proceso de dividir el texto en tokens. Nos referiremos a palabras, tokens o términos indistintamente a lo largo del trabajo.

El tratamiento de texto como palabras individuales nos permite manipular, resumir y visualizar sus características e integrar el procesamiento del lenguaje natural. Para ayudarnos durante el análisis, haremos uso de herramientas incluidas en paquetes como dplyr (Wickham y Francois 2016), tidyr (Wickham 2016), ggplot2 (Wickham 2009) o tidytext (Silge y Robinson 2016). Dentro de R-Studio podremos disponer de más librerías que nos facilitarán el trabajo.

Para una mayor homogeneidad en el trabajo se han fijado una serie de parámetros a la hora de presentar las tablas y las figuras:

- Las tablas o salidas de R-Studio contienen un color de fondo negro para resaltar su contenido. Se mostrarán los 10 primeros resultados a modo de ejemplo para evitar imágenes demasiado largas. En alguna ocasión se mostrarán 20
- Las figuras están generadas con la librería ggplot2, con parámetros fijos como la gama de colores y el tema o fondo del gráfico (estilo minimalista). En ocasiones se eliminará la leyenda así como las escalas en favor de una mejor visualización.

2. OBTENCIÓN DE DOCUMENTOS Y PRIMER VISUALIZADO

Para comenzar este trabajo se deben obtener las cartas a tratar. Al ser documento público, pueden ser descargadas desde la página oficial de Berkshire Hathaway, en el apartado ‘letters’. El primer reto que se encuentra es el formato, teniendo parte de ellas en HTML y otras en formato PDF, por lo que se deben utilizar dos códigos diferentes para su obtención. Una vez iniciado R-Studio, lo primero será importar una serie de librerías (detalladas en los Anexos) para la correcta aplicación de las fórmulas. Una vez descargadas las cartas, se procede a combinarlas en un único documento y así dar comienzo con el análisis.

El primer paso que se debe tener en cuenta para el correcto procesamiento de texto es el de conseguir la forma ordenada del mismo. Para ello, se dividió el conjunto de texto en tokens, con ayuda de la función `unnest_tokens()`, del paquete `tidytext`, que permite posicionar cada palabra en una fila. Gracias a esta función, se pudieron aplicar los principios de limpieza de texto como la eliminación de signos de puntuación y la transformación de mayúsculas a minúsculas.

Para terminar con la limpieza del texto se eliminaron aquellas palabras que carecen de utilidad. Se consideran stopwords a las palabras muy repetidas como artículos, preposiciones o algunos verbos que no aportan información esencial para el análisis. Este es un punto crucial si se quiere que los siguientes procesos de manipulación, procesamiento y visualización surtan efecto. Se hace uso de los paquetes `dplyr`, `tidyverse` y `ggplot2` (Figura 1)

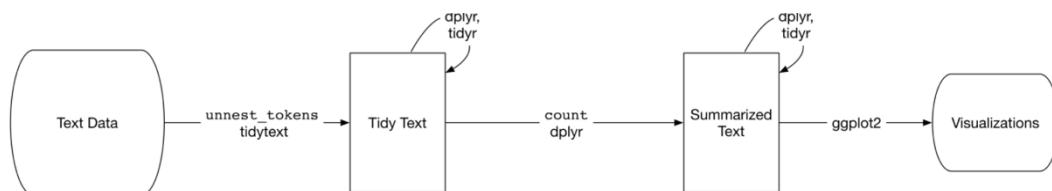


Figura 1: Cadena de transformaciones

Fuente: Julia Silge and David Robinson (2020) “Text Mining with R - A Tidy Approach”

Una vez concluida la limpieza, se realizó un contador del conjunto con la función `count()`. Se observa cuáles son las palabras que más se repiten a lo largo de las 43 cartas así como su frecuencia (tabla 1). Se comprueba que las palabras más repetidas son términos económicos y financieros. Se acompaña de una visualización a modo de gráfico de barras con su frecuencia (figura 2).

Tabla 1: Palabras más frecuentes

	word	n		market	819
1	business	2205	10		
2	berkshire	2182	11	tax	810
3	earnings	1955	12	stock	801
4	company	1312	13	time	748
5	million	1250	14	cost	726
6	insurance	1236	15	shareholders	719
7	businesses	1048	16	charlie	715
8	billion	889	17	capital	699
9	companies	874	18	investment	674
10	market	819	19	shares	672
			20	share	660

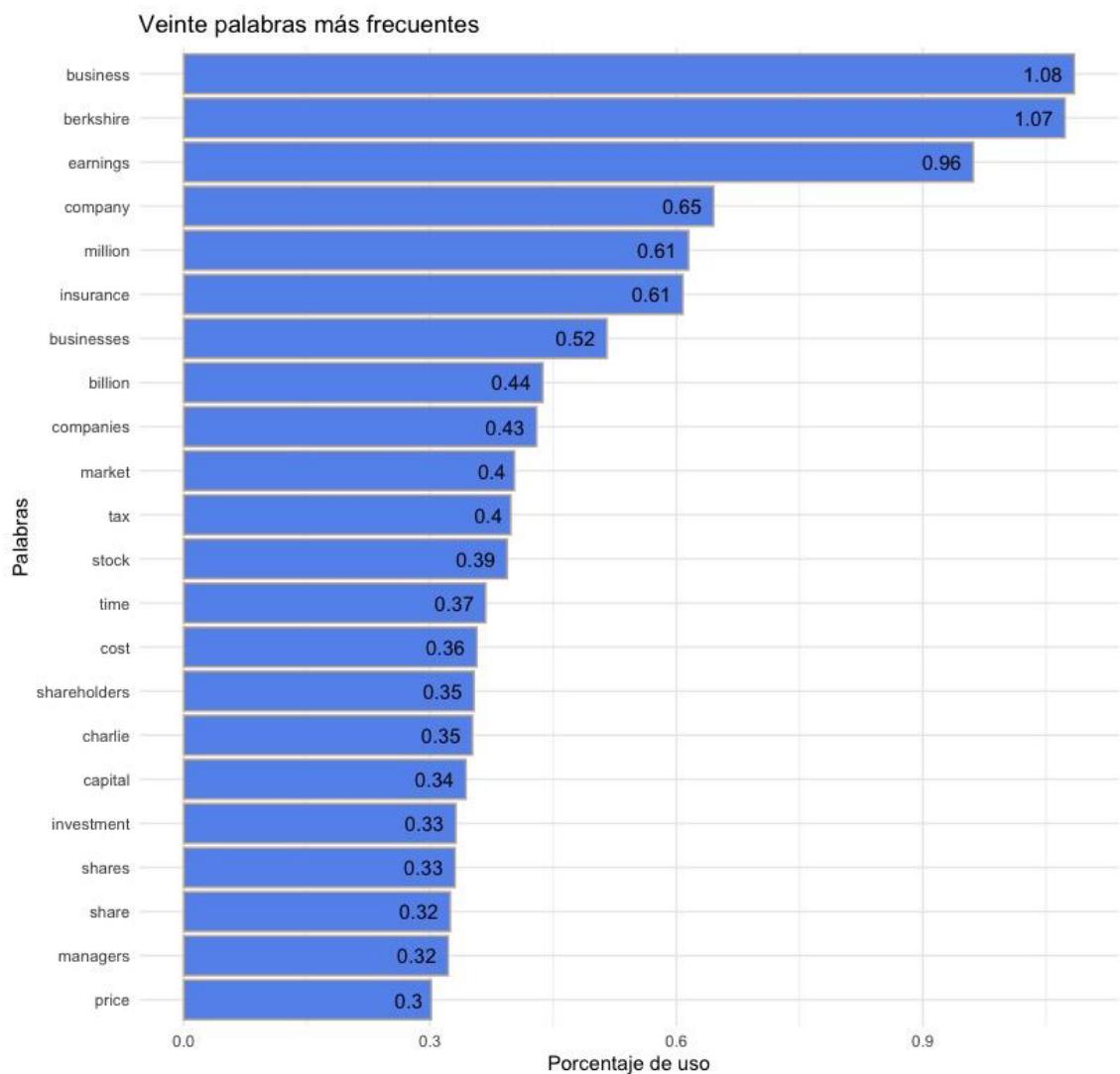


Figura 2: Palabras más frecuentes en %

Aquí se encontró un primer dato interesante. A pesar de que el término ‘business’ es el más repetido con una frecuencia de 2205, solo representa un 1.08% del total de palabras. Este ranking es seguido por ‘berkshire’ (1.07%) y ‘earnings’ (0.98%). Esto permite ver la cantidad tan variada de términos presentes en las cartas.

Para concluir este capítulo, se muestra una nube de palabras donde se representaron las 100 y las 200 palabras con mayor frecuencia (figuras 3 y 4). Una nube de palabras es una representación visual de un documento que nos permite, de un vistazo, conocer su contenido. El tamaño de cada palabra viene determinado por su frecuencia. Nótese que algunas palabras de la primera figura, por su longitud en número de caracteres, han sido suprimidas en favor de la visualización grupal.



Figura 3: Wordcloud de 100 palabras



Figura 4: Wordcloud de 200 palabras

3. MÁS ALLÁ DE LAS PALABRAS

El objetivo de este capítulo ha sido la realización de una manipulación estratégica de las palabras presentes en el conjunto de cartas para tratar de conocer más sobre ellas. Se comenzó por identificar en qué años se encuentran las palabras más repetidas aplicando la función count(), indicando un orden decreciente en frecuencia (tabla 2).

Tabla 2: Palabras más repetidas con sus respectivos años

> words_by_year	year	word	n	10	1989	business	85
	1 2014	berkshire	203	11	1987	business	80
	2 1985	business	112	12	2014	berkshire's	78
	3 1983	business	97	13	2014	earnings	78
	4 1984	business	96	14	2017	berkshire	78
	5 2014	business	92	15	2006	berkshire	77
	6 1990	business	90	16	2000	business	76
	7 2015	berkshire	90	17	1986	business	72
	8 1980	earnings	87	18	2012	berkshire	71
	9 2016	berkshire	86	19	2014	businesses	70
	10 1989	business	85	20	2013	berkshire	69

Se puede observar como estos términos más frecuentes se encuentran distribuidos a lo largo de la mayoría de las cartas.

Se realizó también una comparativa entre la primera carta (1977) y la última (2019) esperando encontrar algún tipo detalle significativo por los 42 años de diferencia. Para ello se aplicaron las funciones filter(), que permite identificar un valor por el cual filtrar el contenido y arrange(), que permite manipular la forma final de nuestra salida y el orden a seguir por nuestras variables. Se obtienen así el top 10 de palabras más repetidas en cada año (tabla 3).

Tabla 3: Palabras más repetidas en 1977 y 2019

year	word	n	year	word	n
1 1977	million	31	1 2019	berkshire	57
2 1977	insurance	30	2 2019	earnings	39
3 1977	earnings	26	3 2019	billion	32
4 1977	company	25	4 2019	company	29
5 1977	capital	18	5 2019	berkshire's	28
6 1977	business	16	6 2019	business	25
7 1977	companies	15	7 2019	companies	24
8 1977	operation	13	8 2019	shares	22
9 1977	national	12	9 2019	insurance	20
10 1977	operating	12	10 2019	stock	19

No se destaca ningún dato más allá de la diferencia en frecuencia total de palabras por cada carta (1375 vs. 3013)

3.1 ¿Qué años contienen las palabras más repetidas?

Para conocer cómo se distribuyen las palabras más frecuentes a lo largo de los años, se fijaron 3 índices basados en la frecuencia de término. El primero consta únicamente de las palabras que superen una frecuencia mayor a 50 (figura 5), el segundo con 60 (figura 6) y el tercero aquellos términos que superen la cifra de 70 (figura 7).

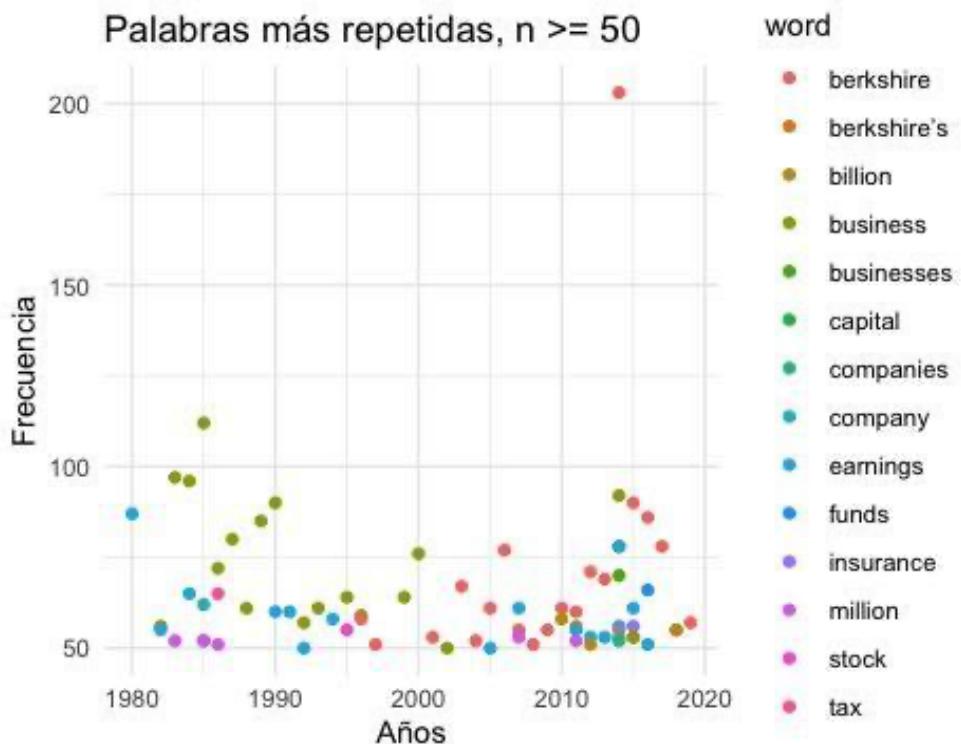


Figura 5: Palabras más repetidas distribuidas en años (n mayor a 50)

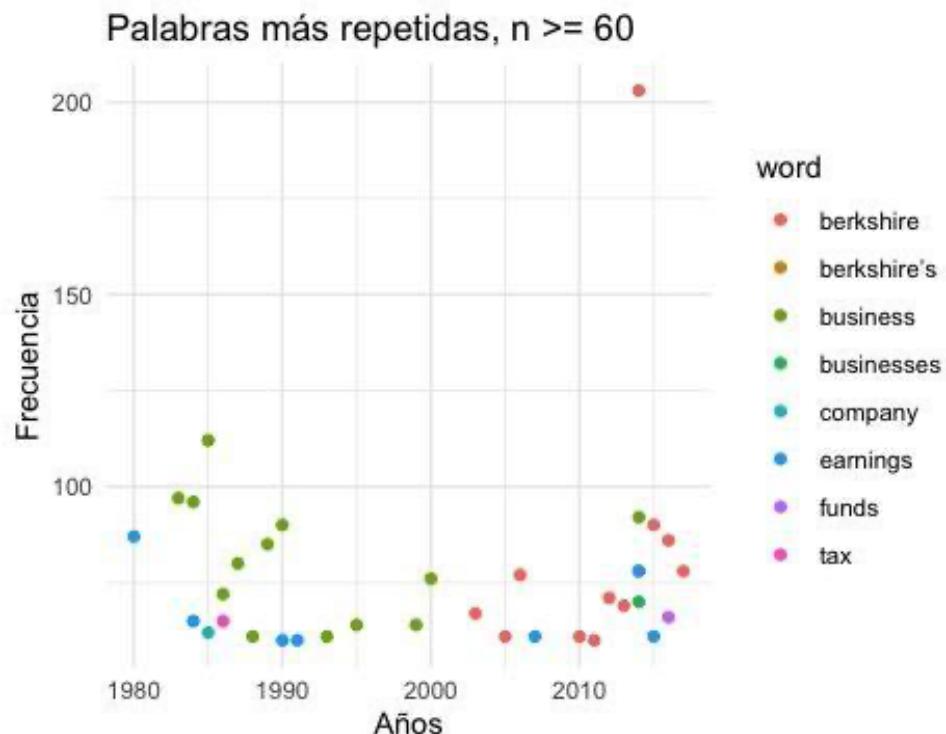


Figura 6: Palabras más repetidas distribuidas en años (n mayor a 60)

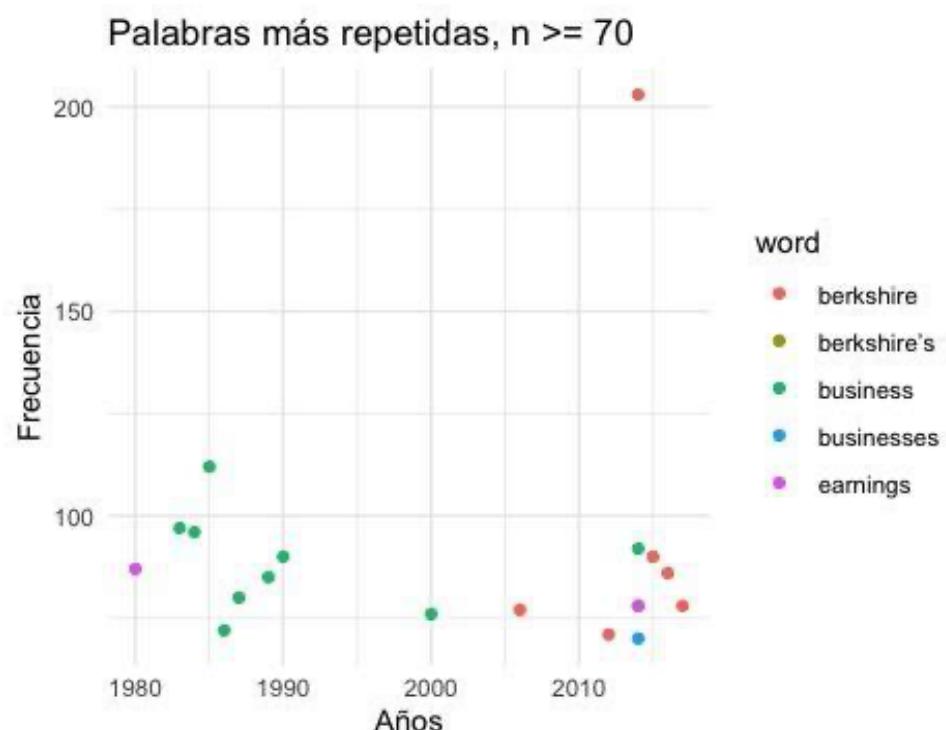


Figura 7: Palabras más repetidas distribuidas en años (n mayor a 70)

Se comprueba cómo estas palabras se van repitiendo a lo largo de los años de una forma normal. Destaca el dato un tanto atípico de 2014 en el que se repite 204 veces la palabra ‘berkshire’.

3.2 Frecuencias totales por cada año

Para obtener el sumatorio total de palabras por cada año, se hizo uso de la función `group_by()` para indicar el valor de agrupación y `summarize()` para generar el resumen creando la nueva variable `sum.words`. Se aplicó la función `View()` para abrir una nueva pestaña en R-Studio y visualizar el resultado con una mayor claridad.

La tabla 4 muestra la agrupación del año y su frecuencia. Se aplicó un orden creciente en años, pero también se podría ordenar por número de palabras, resultando así el top 3 en frecuencia los años 2014, 2015 y 2016.

Tabla 4: Sumatorio de palabras por cada año

	year	sum.words	20	1996	4515
1	1977	1375	21	1997	4490
2	1978	1806	22	1998	4375
3	1979	2707	23	1999	4747
4	1980	3170	24	2000	5238
5	1981	2843	25	2001	4844
6	1982	3284	26	2002	5891
7	1983	4582	27	2003	5308
8	1984	4980	28	2004	5809
9	1985	5435	29	2005	5269
10	1986	5553	30	2006	5679
11	1987	4867	31	2007	4808
12	1988	4789	32	2008	5300
13	1989	5758	33	2009	4455
14	1990	6091	34	2010	5726
15	1991	3646	35	2011	5622
16	1992	4334	36	2012	5576
17	1993	4321	37	2013	5318
18	1994	3665	38	2014	9532
19	1995	4462	39	2015	6961
20	1996	4515	40	2016	6575
21	1997	4490	41	2017	3647
22	1998	4375	42	2018	2969
			43	2019	3013

Se procede a graficar esta distribución de palabras en forma de línea y puntos (figura 8). Para la realización de este gráfico se hizo uso de la librería `ggplot2`, la cual permite trabajar por capas, agregando las líneas de código necesarias para alcanzar los objetivos. En este caso, se aplicó la capa ‘`geom_line`’ para trazar las líneas, seguida de ‘`geom_point`’ para graficar los puntos por encima. A continuación se muestra la tendencia (figura 9).

Palabras totales por año

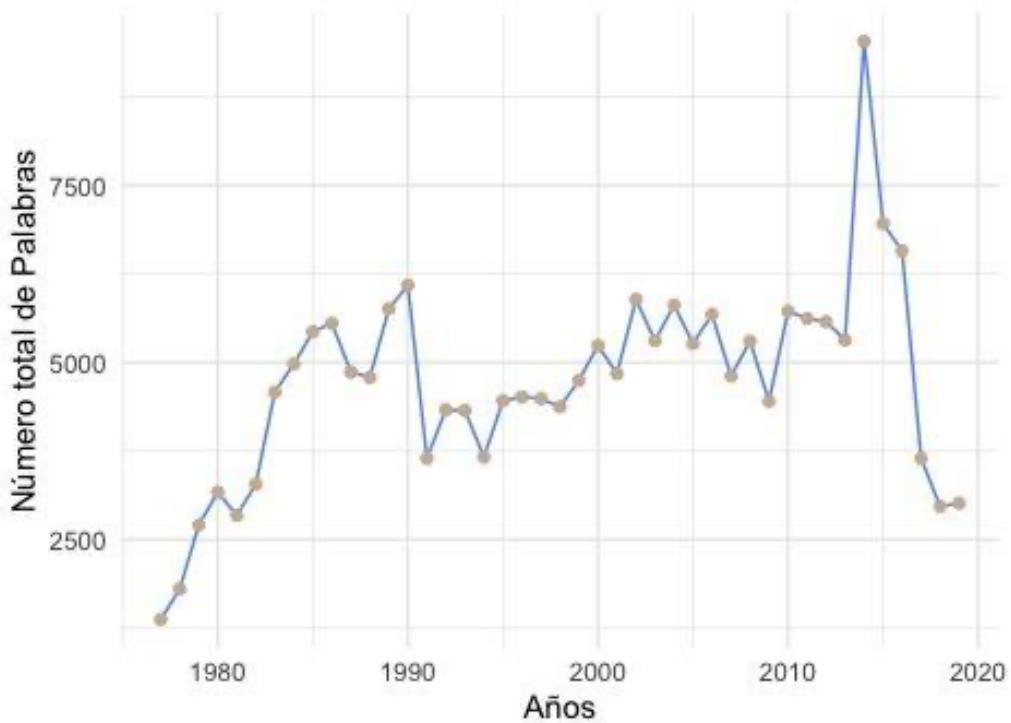


Figura 8: Palabras totales por años. Gráfico de líneas

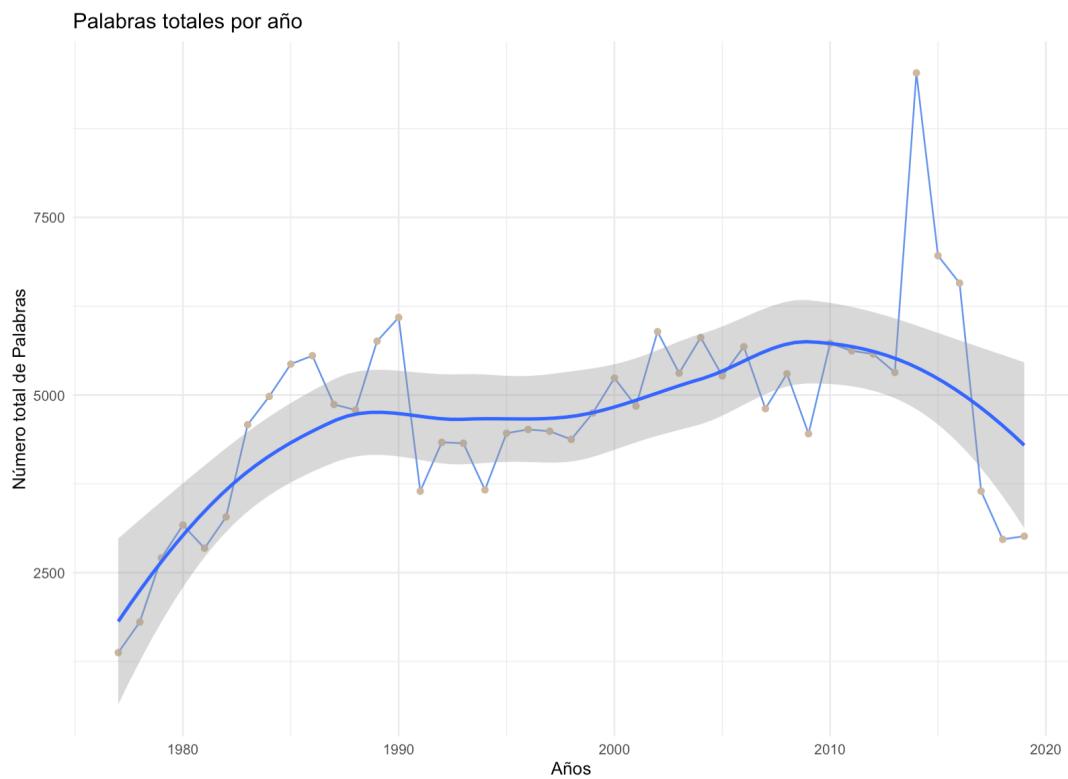


Figura 9: Palabras totales por años. Gráfico de tendencia

Se puede observar cómo, a pesar de la baja cantidad de datos para hablar de una clara tendencia, hasta 2014 la pendiente era positiva (tendencia creciente) para el número de palabras en cada carta. A partir de esta fecha ha ido disminuyendo.

3.3 Palabras más repetidas en cada año

En este punto la intención es obtener la frecuencia de palabras dispuesta en cada año. Para ello, primero se fijó el año y después se buscó la repetición de palabras, obteniendo así los términos más repetida en cada año (tabla 5). Aparecen las palabras ‘million’, ‘earnings’, ‘business’, ‘berkshire’ y ‘billion’ como las más repetidas en el conjunto de los años. Estos términos económicos y financieros caracterizan el contenido de las cartas.

Tabla 5: La palabra más repetida en cada año

#	year	word	n	#	22	1998	earnings	49
1	1977	million	31		23	1999	business	64
2	1978	earnings	34		24	2000	business	76
3	1979	earnings	38		25	2001	berkshire	53
4	1980	earnings	87		26	2002	business	50
5	1981	earnings	48		27	2003	berkshire	67
6	1982	business	56		28	2004	berkshire	52
7	1983	business	97		29	2005	berkshire	61
8	1984	business	96		30	2006	berkshire	77
9	1985	business	112		31	2007	earnings	61
10	1986	business	72		32	2008	berkshire	51
11	1987	business	80		33	2009	berkshire	55
12	1988	business	61		34	2010	berkshire	61
13	1989	business	85		35	2011	berkshire	60
14	1990	business	90		36	2012	berkshire	71
15	1991	earnings	60		37	2013	berkshire	69
16	1992	business	57		38	2014	berkshire	203
17	1993	business	61		39	2015	berkshire	90
18	1994	earnings	58		40	2016	berkshire	86
19	1995	business	64		41	2017	berkshire	78
20	1996	business	59		42	2018	berkshire	55
21	1997	berkshire	51		43	2018	billion	55
					44	2019	berkshire	57

Se profundiza hacia una visualización de las 5 palabras más repetidas en cada año (figura 10) ampliando de esta forma el abanico de términos presentes en las cartas del señor Buffett. Aquí se pudieron observar, gracias a la repetición de puntos del mismo color, como año tras año los términos económicos ocupan los primeros puestos del ranking. Se modificó la escala del eje x, mostrándose todo el conjunto de valores (años) para una correcta interpretación.

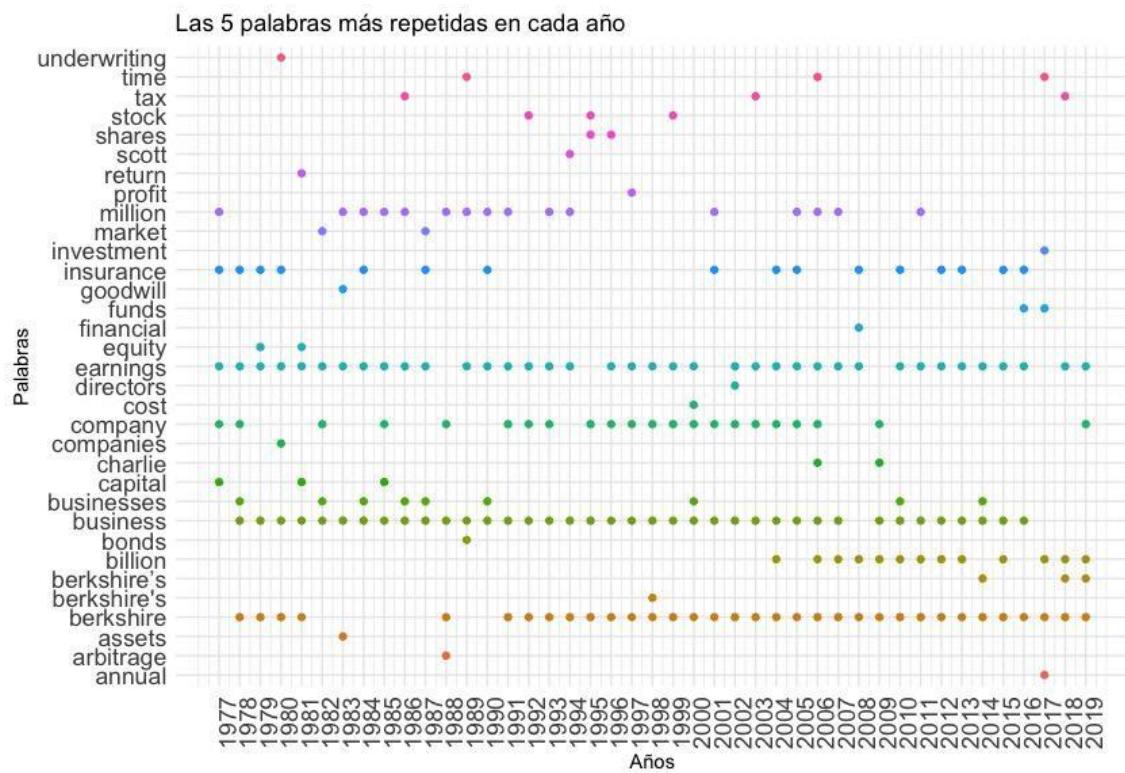


Figura 10: Top 5 palabras más repetidas en cada año

Gracias a esta clasificación, aparecen nuevos términos como ‘equity’ (1979 y 1981), ‘underwriting’ (1980), ‘tax’ (1986 y 2003) o ‘assets’ (1983). Se observa cómo a medida que se profundiza en cada carta, se encuentran más detalles sobre el contenido, pudiendo de esta forma, hacerse una idea del tema general a grandes rasgos.

4. ANÁLISIS DE SENTIMIENTO A TRAVÉS DE LOS AÑOS

En el capítulo anterior se exploró el contenido de las cartas, centrándose en la frecuencia de las palabras y su visualización. A partir de ahora se plantea el análisis de sentimiento. En la mayoría de ocasiones, se puede utilizar el propio nivel de comprensión lectora para interpretar un texto. Se habla de esta manera de textos positivos o negativos, optimistas o pesimistas. La finalidad de este capítulo es ayudar a realizar una interpretación lo más veraz posible del sentimiento a través de los años. Gracias al paquete tidytext se accede a cuatro léxicos;

- AFINN (Finn Årup Nielsen). Asigna puntuación para cada palabra (-4 ; 5).
- Loughran (Loughran, T. and McDonald, B.) Clasifica en 6 grupos.
- Bing (Bing Liu y colaboradores) Clasifica entre positivo o negativo.
- Nrc (Saif Mohammad y Peter Turney). Clasifica en 8 grupos .

Se comienza con un resumen general del contenido de estos léxicos. Para ello se empleó la función `get_sentiments()` seguido del léxico al que se quiere acceder (tabla 6)

Tabla 6: Léxicos

> get_sentiments("afinn")		> get_sentiments("loughran")	
# A tibble: 2,477 x 2		# A tibble: 4,150 x 2	
word	value	word	sentiment
<chr>	<dbl>	<chr>	<chr>
1 abandon	-2	1 abandon	negative
2 abandoned	-2	2 abandoned	negative
3 abandons	-2	3 abandoning	negative
4 abducted	-2	4 abandonment	negative
5 abduction	-2	5 abandonments	negative
6 abductions	-2	6 abandons	negative
7 abhor	-3	7 abdicated	negative
8 abhorred	-3	8 abdicates	negative
9 abhorrent	-3	9 abdicating	negative
10 abhors	-3	10 abdication	negative
# ... with 2,467 more rows		# ... with 4,140 more rows	
> get_sentiments("nrc")		> get_sentiments("bing")	
# A tibble: 13,901 x 2		# A tibble: 6,786 x 2	
word	sentiment	word	sentiment
<chr>	<chr>	<chr>	<chr>
1 abacus	trust	1 2-faces	negative
2 abandon	fear	2 abnormal	negative
3 abandon	negative	3 abolish	negative
4 abandon	sadness	4 abominable	negative
5 abandoned	anger	5 abominably	negative
6 abandoned	fear	6 abominate	negative
7 abandoned	negative	7 abomination	negative
8 abandoned	sadness	8 abort	negative
9 abandonment	anger	9 aborted	negative
10 abandonment	fear	10 aborts	negative
# ... with 13,891 more rows		# ... with 6,776 more rows	

Se puede observar cómo el léxico Afinn contiene 2477 palabras, Loughran 5150, Bing 6787 y Nrc 13901 palabras. Se debe tener en cuenta una serie de puntos antes de comenzar nuestro análisis:

1. Determinadas palabras no fueron clasificadas al poseer un sentimiento neutro y por lo tanto podrían llevar a conclusiones incorrectas.
2. Únicamente se analizaron las palabras como unidades individuales, es decir, no se tuvieron en cuenta ni la palabra anterior ni la posterior. Por ejemplo en la siguiente oración: “Mi hermano no es bueno”, el sentimiento de ‘bueno’ será positivo a pesar de que en su conjunto podría ser clasificado como negativo.
3. Se realizó un análisis literal de cada palabra, no se tuvo en cuenta ninguna interpretación de estructuras retóricas.

Para la correcta combinación de cada léxico con las cartas se aplicó la función “inner_join()” obteniendo así el resultado deseado.

4.1 Afinn

Una vez aplicado este léxico a las cartas, se obtuvieron un total de 1238 palabras clasificadas (tabla 5.1.1) de las 2477 disponibles en el propio léxico. Se presenta una calificación entre ‘-4’ y ‘-1’ para los sentimientos negativos y ‘1’ a ‘5’ para indicar el sentimiento positivo de cada palabra. Se agrega la columna ‘contribution’ resultado de multiplicar la frecuencia de cada palabra por la puntuación que arroja este léxico (tabla 7). A continuación se indican las 30 palabras (positivas y negativas) que más contribuyen al sentimiento (figura 11).

Tabla 7: Palabras clasificadas - Afinn

letters_sentiments_afinn_word		
word	occurrences	contribution
<chr>	<int>	<dbl>
1 abandon	5	-10
2 abandoned	4	-8
3 abhor	2	-6
4 abilities	13	26
5 ability	99	198
6 aboard	3	3
7 absentee	6	-6
8 absolves	1	2
9 absorbed	3	3
10 abuse	2	-6
# ... with 1.228 more rows		

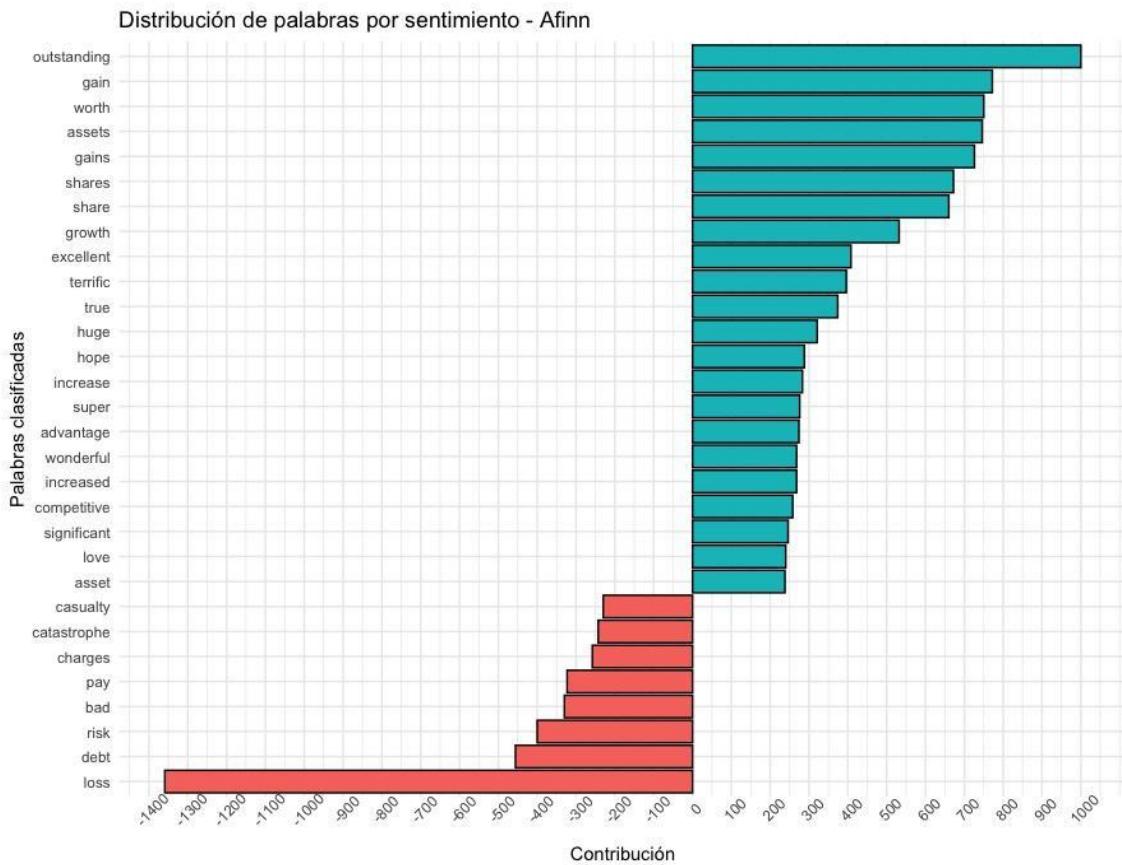


Figura 11: Distribución de palabras por sentimiento - Afinn (Top 30)

Podemos notar una clasificación interesante sobre el sentimiento que pretende plasmar el señor Buffett. Se observa que, de las 30 palabras que más contribuyen a la clasificación, 22 son positivas y 8 negativas. Más allá de las palabras esperadas, debido al componente financiero y económico, se reconocen otras no tan comunes. En el lado positivo: 'sobresaliente' o 'extraordinario' y en el lado negativo: 'inusual', 'difícil' o 'malo'. Se verifica cómo el lenguaje empleado anima más a generar positivismo y esperanza.

Se examinaron cuales fueron las palabras con mayor puntuación positiva (sentiment = +5) en cada carta, obteniendo como resultado “outstanding” (excepcional) en 20 de las 22 cartas con esta calificación (Tabla 8).

Tabla 8: Palabras Positivas clasificadas en cada año - Afinn

▼	year	word	sentiment	occurrences	11	1998	outstanding	5	5
1	1979	outstanding	5	6	12	1999	outstanding	5	6
2	1984	outstanding	5	7	13	2000	outstanding	5	9
3	1986	outstanding	5	7	14	2002	outstanding	5	7
4	1988	superb	5	5	15	2004	outstanding	5	5
5	1989	outstanding	5	5	16	2005	outstanding	5	8
6	1989	superb	5	6	17	2006	outstanding	5	9
7	1992	outstanding	5	7	18	2008	outstanding	5	6
8	1995	outstanding	5	9	19	2009	outstanding	5	5
9	1996	outstanding	5	9	20	2011	outstanding	5	8
10	1997	outstanding	5	5	21	2012	outstanding	5	9
					22	2014	outstanding	5	10

Una vez aplicado el léxico y realizado el primer análisis, se agregó una nueva variable “score”, que hace referencia a la puntuación promedio del sentimiento para cada año (Figura 12). Se ordenó cronológicamente cada carta seguida de la puntuación obtenida. Se apreció que la carta perteneciente al año 2001 arroja un “score” negativo, pudiendo ser causado por el colapso de las denominadas “punto-com” ligado a los atentados del 11-S. También se notó la carta perteneciente al año 2008 por estar muy próxima al valor neutro. La Gran Recesión y otros eventos relacionados durante este período influyeron en la crisis financiera de 2007-2008. La crisis hipotecaria también jugó un papel muy importante en esta época.

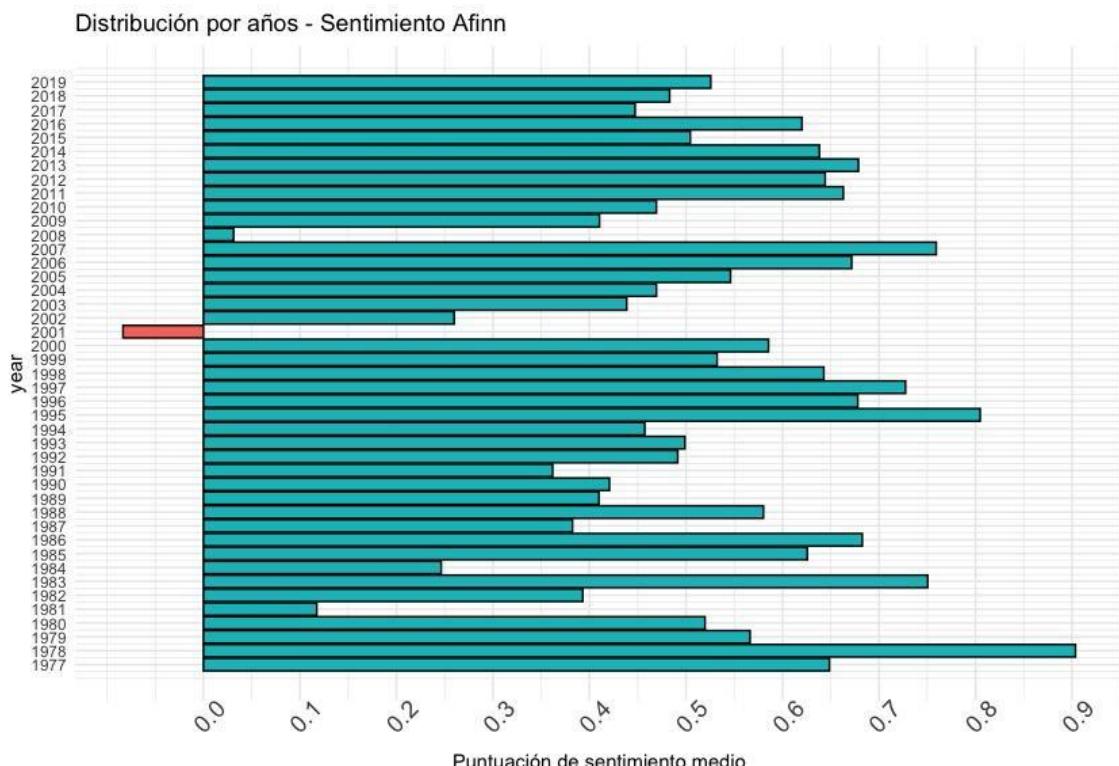


Figura 12: Distribución por años - Sentimiento Afinn (barras)

A continuación se muestran los mismos datos en formato de puntos y añadiendo una capa de tendencia (Figura 13). Se reconoce como esta se mueve en el intervalo (0.45 ; 0.625) aproximadamente, mostrando el optimismo presente en las cartas. Desde el año 2003 estaríamos en una tendencia positiva.

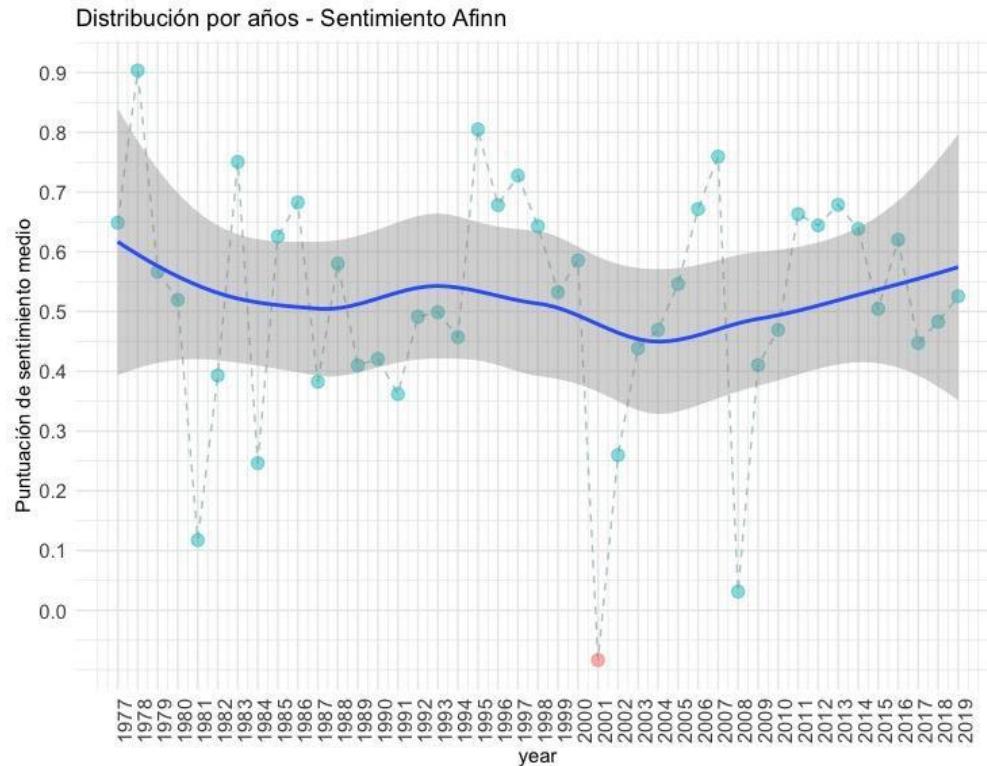


Figura 13: Distribución por años - Sentimiento Afinn (puntos y tendencia)

4.2 Loughran

En esta ocasión se aplicó un nuevo léxico que nace en base a análisis de informes financieros. El diccionario Loughran divide las palabras en seis sentimientos: ‘positive’, ‘negative’, ‘litigious’, ‘uncertainty’, ‘constraining’ y ‘superfluous’. Se obtuvieron un total de 1593 palabras diferentes clasificadas (tabla 9). Se presenta también una división por sentimiento general (contenido en el propio léxico) seguido de la aplicación a nuestros documentos (tabla 10).

Tabla 9: Palabras registradas en Loughran

	word	n	sentiment
	<chr>	<int>	<chr>
1	abandon	5	negative
2	abandoned	4	negative
3	abandoning	3	negative
4	abandonment	1	negative
5	aberration	1	negative
6	aberrational	6	negative
7	aberrations	5	negative
8	abnormal	1	negative
9	abnormalities	1	negative
10	abnormality	1	negative
	# ... with 1,583 more rows		

Tabla 10: Contador de palabras clasificadas en Loughran (General vs Aplicado)

> loughran_sentiment		> letters_sentiments_loughran_count	
sentiment	n	sentiment	n
1 constraining	184	1 constraining	985
2 litigious	904	2 litigious	1068
3 negative	2355	3 negative	7194
4 positive	354	4 positive	5235
5 superfluous	56	5 superfluous	37
6 uncertainty	297	6 uncertainty	1864

Se profundiza hacia una visualización de las 5 palabras que más contribuyen a cada sentimiento (figura 14). Se reconoce ‘loss’ encabezando el sentimiento ‘negative’ y ‘gain’ para el ‘positive’, pero también se observa cómo ‘risk’ encabeza el ‘uncertainly’ y ‘contracts’ el ‘litigious’.



Figura 14: Clasificación de palabras por sentimiento - Loughran

Se aplicaron estos sentimientos para visualizar la clasificación para cada carta a lo largo de los años (figura 15). Nótese la cantidad de palabras clasificadas como ‘negative’ (color verde) que aparecen en cada año. Para clarificar esta situación se puede consultar la tabla 10, hallando así como la gran mayoría de palabras clasificadas (2355 de 4150) pertenecen a este sentimiento.

Clasificación Loughran por año

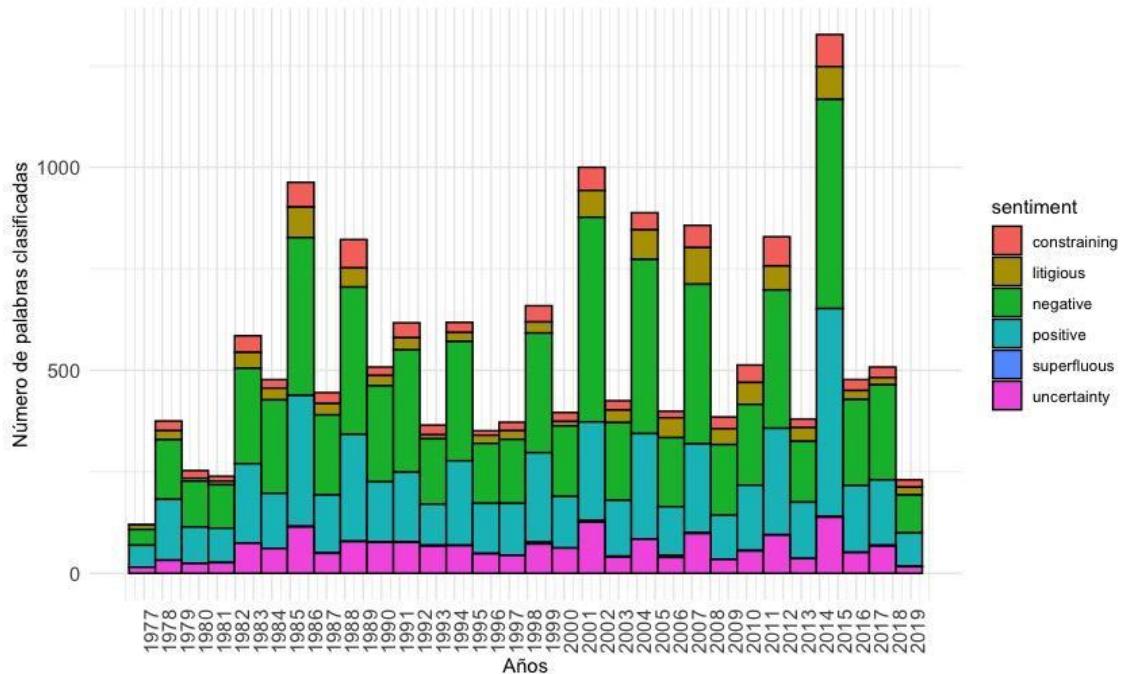


Figura 15: Distribución por años - Sentimiento Loughran

Con motivo de una mayor comprensión se dividió el contenido en 6 diferentes visualizaciones, una por cada sentimiento. Para ello se agregó la capa de ‘facets()’ en ggplot2 y se permitió que las escalas actusasen libremente (figura 16).

Se distinguen los contrastes, en frecuencia de palabras clasificadas, para el sentimiento ‘superfluous’ con un máximo en 5, frente a ‘positive’ con un valor que supera la cifra de 500.

Se refleja a continuación la división independiente por cada año pero con escala fija para observar visualmente los contrastes entre cada carta (figura 17).

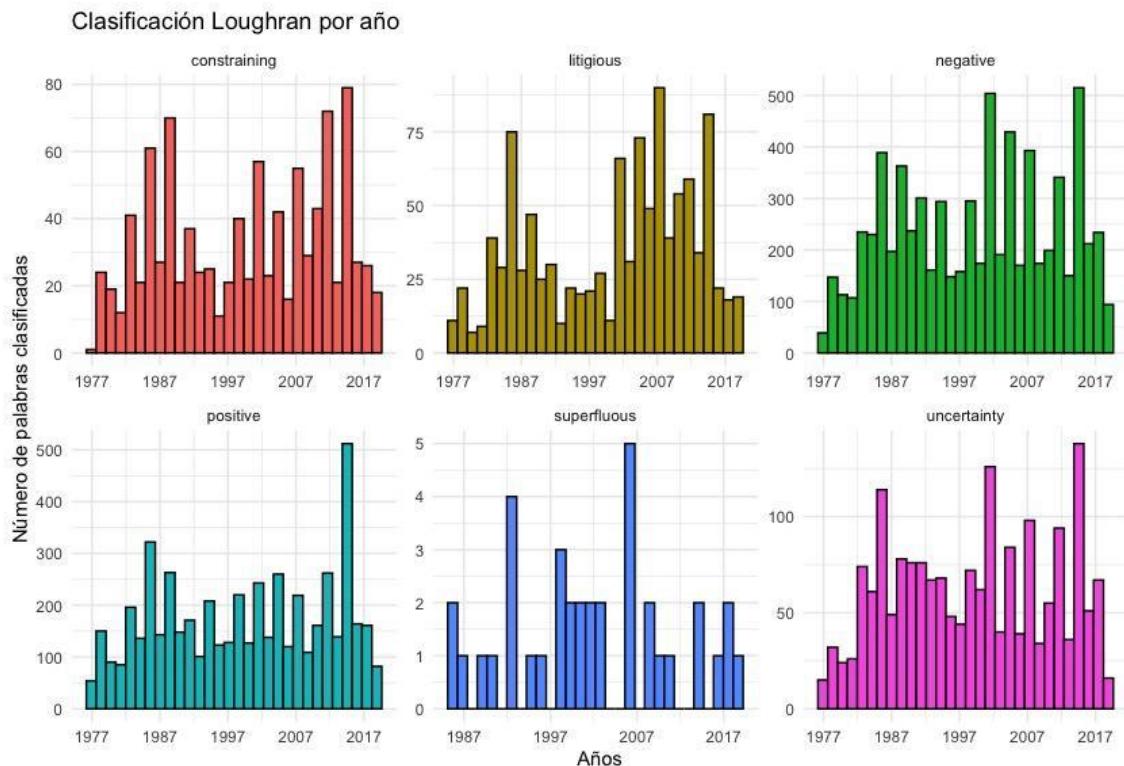


Figura 16: Distribución por años - Sentimiento Loughran (dividido por sentimiento)

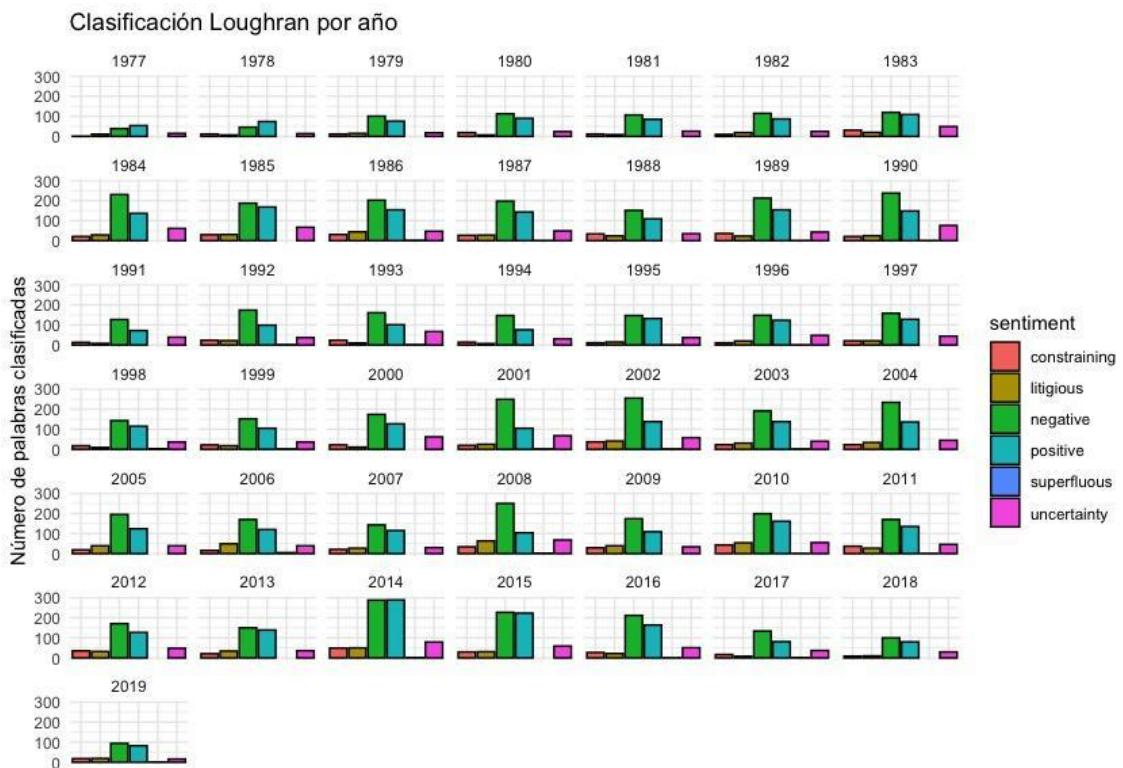


Figura 17: Distribución por años - Sentimiento Loughran (dividido por años)

Se concluye este apartado dedicado al sentimiento Loughran destacando su gran potencial, permitiendo obtener una valoración más allá de la ya acostumbrada positiva o negativa.

4.3 Nrc

Se continúa con otro léxico, esta vez añadiendo más sentimientos a la clasificación: ‘positive’, ‘negative’, ‘anger’, ‘anticipation’, ‘disgust’, ‘fear’, ‘joy’, ‘sadness’, ‘surprise’ y ‘trust’. Se comprueba como 6253 palabras diferentes fueron clasificadas con éxito (tabla 11)

Tabla 11: Clasificación en Nrc

> letters_sentiments_nrc_word		
# A tibble: 6,253 x 3		
# Groups: sentiment [10]		
word	n	sentiment
<chr>	<int>	<chr>
1 abandon	5	fear
2 abandon	5	negative
3 abandon	5	sadness
4 abandoned	4	anger
5 abandoned	4	fear
6 abandoned	4	negative
7 abandoned	4	sadness
8 abandonment	1	anger
9 abandonment	1	fear
10 abandonment	1	negative
# ... with 6,243 more rows		

Como en anteriores ocasiones, se procedió a solicitar en R-Studio un contador del número de palabras agrupadas por cada sentimiento (tabla 12). Se expone como de las 92.755 palabras clasificadas, el sentimiento ‘positive’ encabeza el ranking de frecuencia, por lo que se verá significativamente reflejado en las cartas.

Tabla 12: Contador de palabras registradas en Nrc (General vs Aplicado)

> nrc_sentiment		> letters_sentiments_nrc_count	
# A tibble: 10 x 2		# A tibble: 10 x 2	
# Groups: sentiment [10]		# Groups: sentiment [10]	
sentiment	n	sentiment	n
<chr>	<int>	<chr>	<int>
1 anger	1247	1 anger	4264
2 anticipation	839	2 anticipation	10739
3 disgust	1058	3 disgust	2085
4 fear	1476	4 fear	5817
5 joy	689	5 joy	8051
6 negative	3324	6 negative	11760
7 positive	2312	7 positive	25236
8 sadness	1191	8 sadness	6028
9 surprise	534	9 surprise	3635
10 trust	1231	10 trust	15160

Se inspeccionó cuales fueron las 5 palabras que más contribuyeron a cada sentimiento (figura 18).

Se destaca cómo ‘tax’ (impuestos) encabeza el sentimiento ‘negative’ y ‘share’ (acciones) el ‘positive’. También se observa cómo, en ocasiones, un mismo término puede tener diferentes connotaciones. La palabra ‘money’ es clasificada tanto para ‘anger’ como para ‘surprise’.

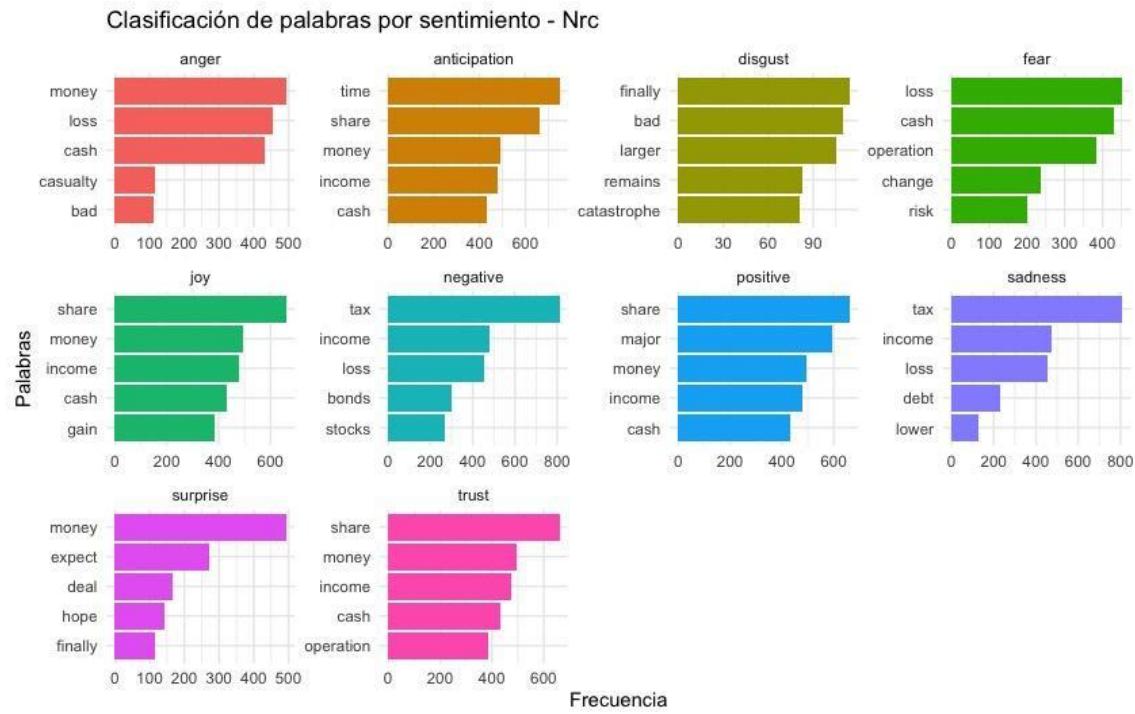


Figura 18: Clasificación de palabras por sentimiento - Nrc

Se aplicaron estos sentimientos con objeto de visualizar la clasificación para cada carta a lo largo de los años (figura 19). Se acompaña de la división por sentimiento (figura 20). Nótese que se elimina la información del eje x para una mejor visualización.

Se verifica como los sentimientos: ‘anger’, ‘disgust’, ‘fear’, ‘surprise’ o ‘sadness’ poseen baja representación frente a ‘positive’ o ‘trust’.

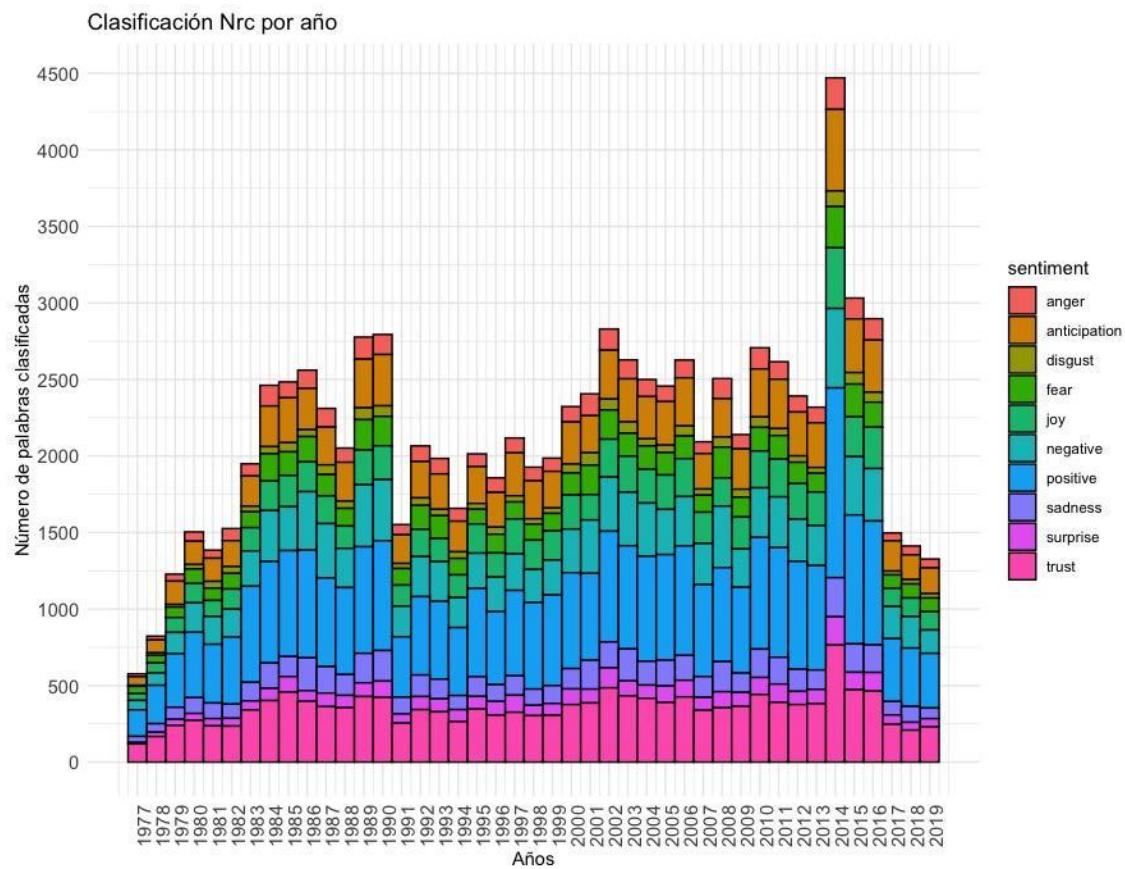


Figura 19: Distribución por años - Sentimiento Nrc

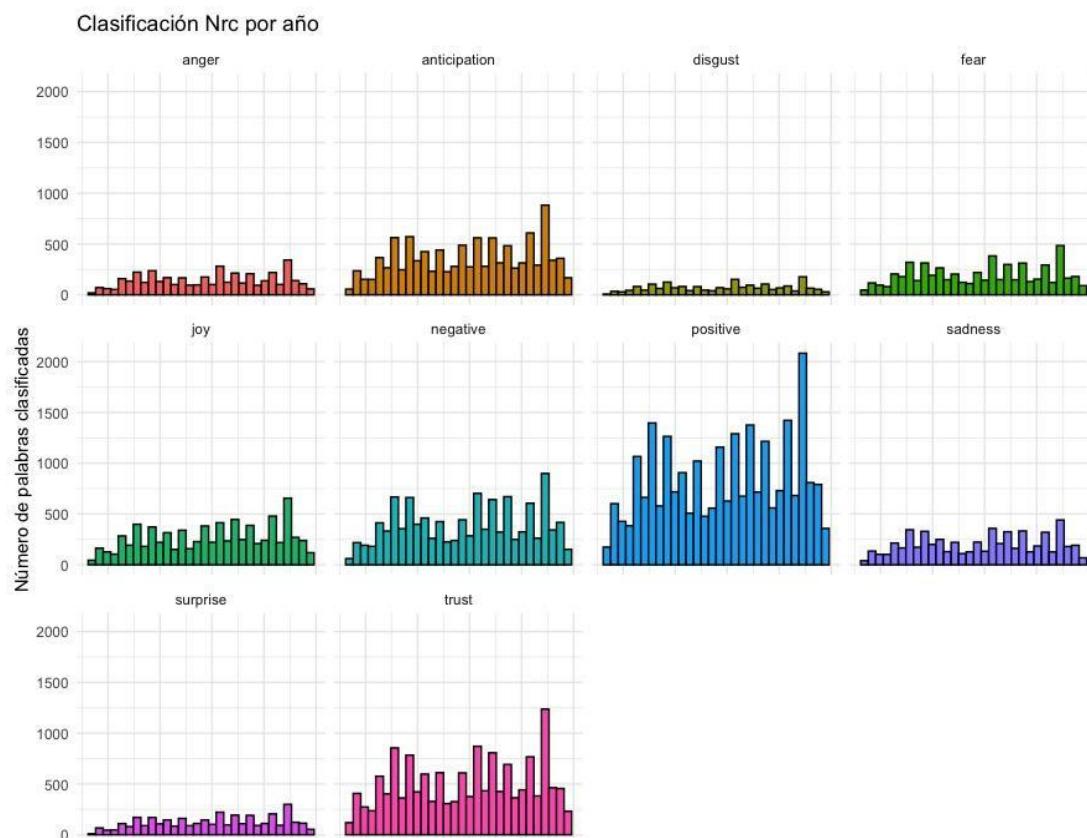


Figura 20: Distribución por años - Sentimiento Nrc (dividido por sentimiento)

Se finaliza este apartado con la realización de una división por cada año, visualizando cómo se distribuyen los 10 sentimientos que posee el léxico “Nrc” a lo largo de las cartas (figura 21).

Se expone una clara relación positiva entre la frecuencia total de cada carta y el total de palabras clasificadas por sentimientos. Ejemplo de esto es el documento perteneciente a 2014.

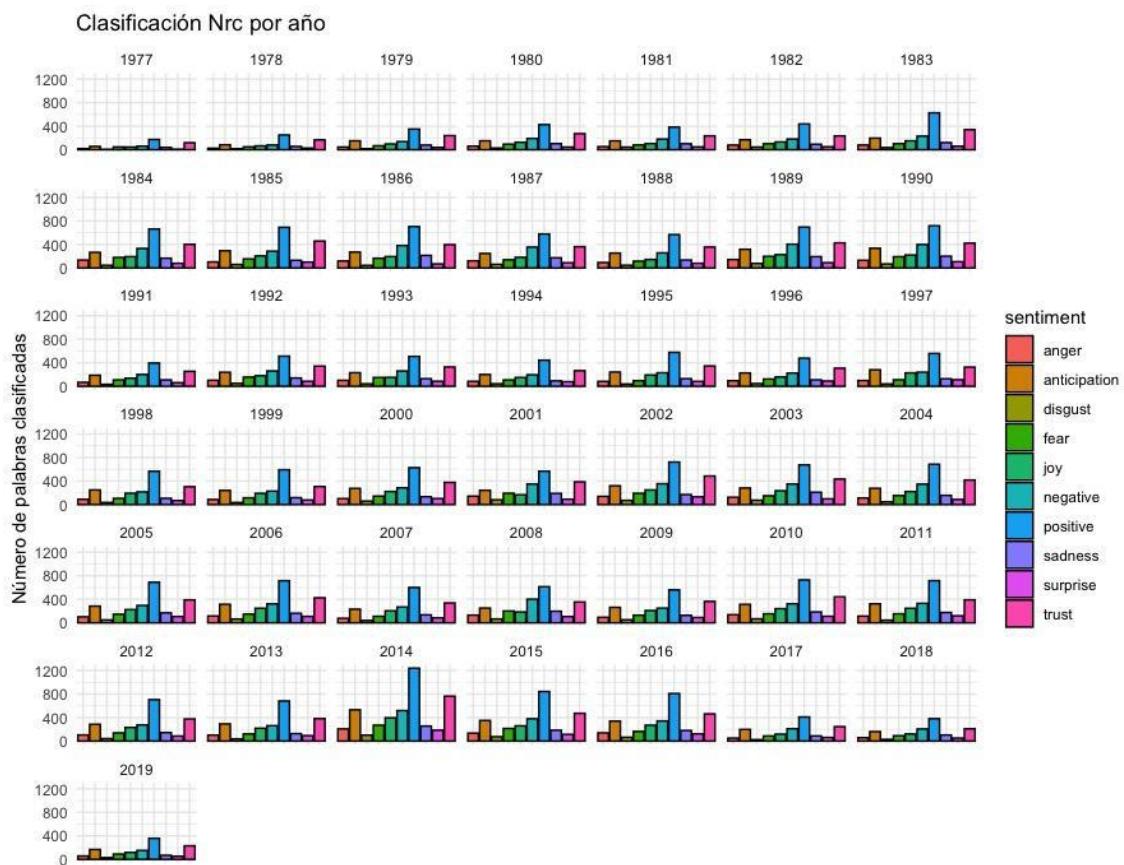


Figura 21: Distribución por años - Sentimiento Nrc

Gracias a este léxico se pudo ampliar el número de sentimientos obtenidos, interpretando así de mejor forma el modo en el que el señor Buffett se dirige a sus accionistas. Se destaca la alta cantidad de palabras clasificadas como “positive” que presenta este léxico frente al de Loughran.

4.4 Bing

El léxico Bing permite una división entre ‘positive’ y ‘negative’. Se aplicó de nuevo la función “get_sentiment()” para incorporar este nuevo léxico a las cartas (tabla 13). Se presenta la cantidad de palabras que contiene para cada sentimiento y también aplicado a nuestros textos. Se registra que en esta ocasión el doble de palabras negativas que positivas, pero una vez aplicado a las cartas esta relación no se cumple. Se verifican 9349 palabras negativas frente a 11902 positivas (tabla 14).

Tabla 13: Clasificación en Bing

# A tibble:	2,370 x 3	
# Groups:	word [2,369]	
word	n	sentiment
<chr>	<int>	<chr>
1 abnormal	1	negative
2 abominable	2	negative
3 abomination	1	negative
4 aborted	1	negative
5 abound	3	positive
6 abounds	1	positive
7 abruptly	1	negative
8 absence	3	negative
9 absentee	6	negative
10 absurd	4	negative
# ... with 2,360 more rows		

Tabla 14: Contador de palabras registradas en Bing (General vs Aplicado)

||
||
||
||
||
||
||
||
||
||
||
||

||
||
||
||
||
||
||
||
||
||
||
||

En esta ocasión se optó por una alternativa de visualización ya utilizada en el capítulo 3. Se extrae una nube de palabras aplicando el filtro de negativo o positivo para mejorar su interpretación (figura 22). Se exponen los términos ‘gain’ o ‘worth’ contribuyendo al sentimiento positivo y ‘loss’ o ‘debt’ para el negativo

Se avanza hacia una división general por años (figura 23). Se aplicó también la capa ‘facets()’ para obtener gráficos independientes por sentimiento (figura 24) y por año (figura 25). Se manifiesta como en los años 1987, 2001 y 2008 los sentimientos negativos superan a los positivos. Se pueden añadir también, por proximidad al valor neutro, los años 1990 y 2002 desde el lado negativo y los años 1981 y 1984 desde el positivo.

Debido a la semejanza que se puede encontrar con el léxico Affin, en esta ocasión se optó por presentar una serie de visualización distintas a las del apartado 4.1, pudiendo así demostrar las diferentes opciones de analizar un mismo sentimiento.



Figura 22: Nube de palabras - Sentimiento Nrc

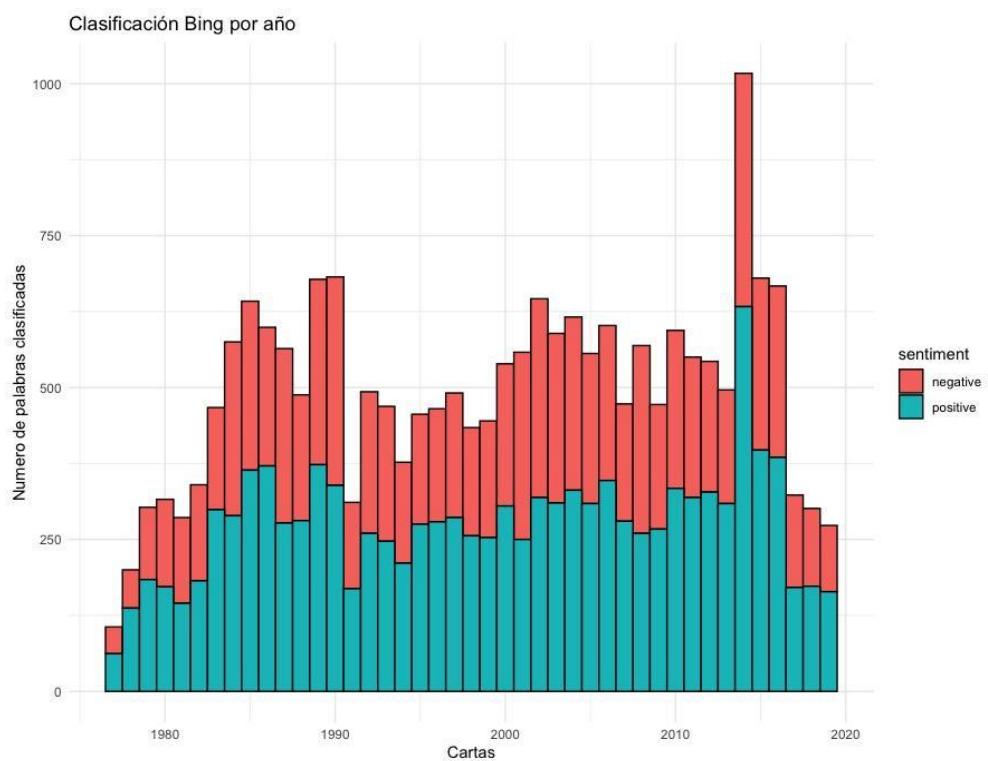


Figura 23: Distribución por años - Sentimiento Nrc

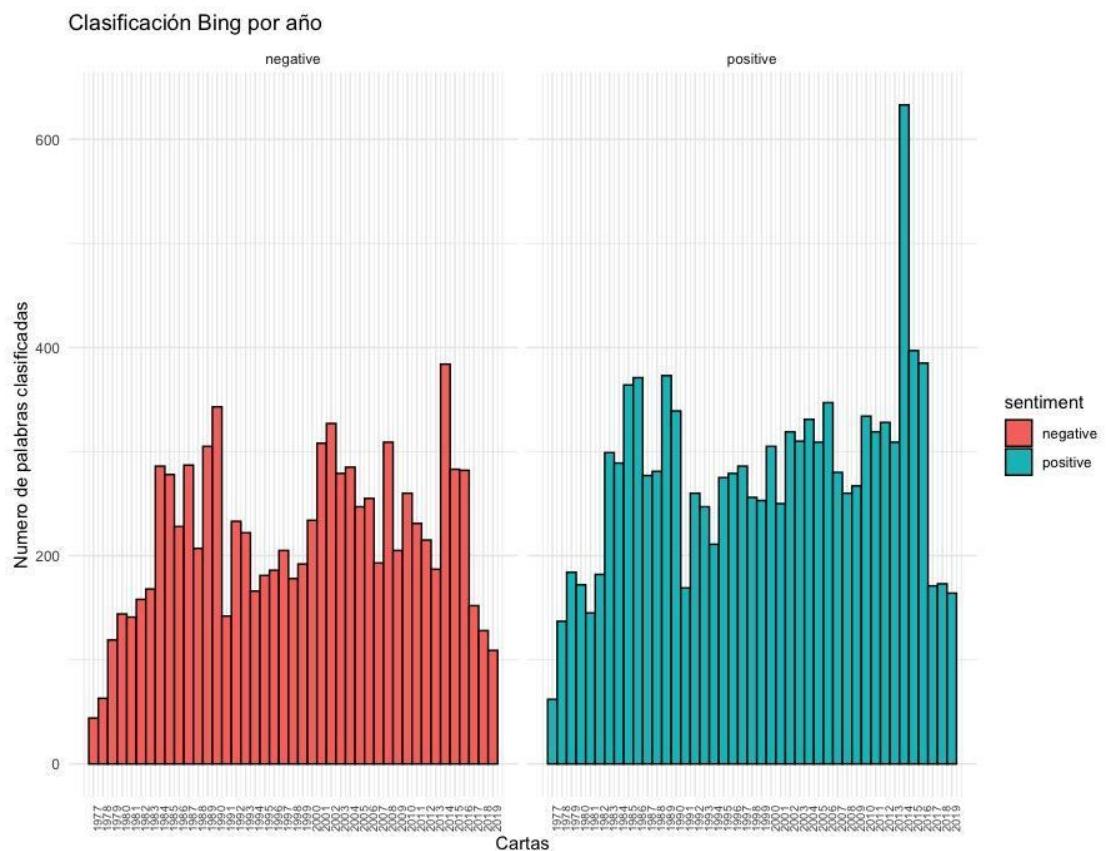


Figura 24: Distribución por años - Sentimiento Nrc (dividido por sentimiento)

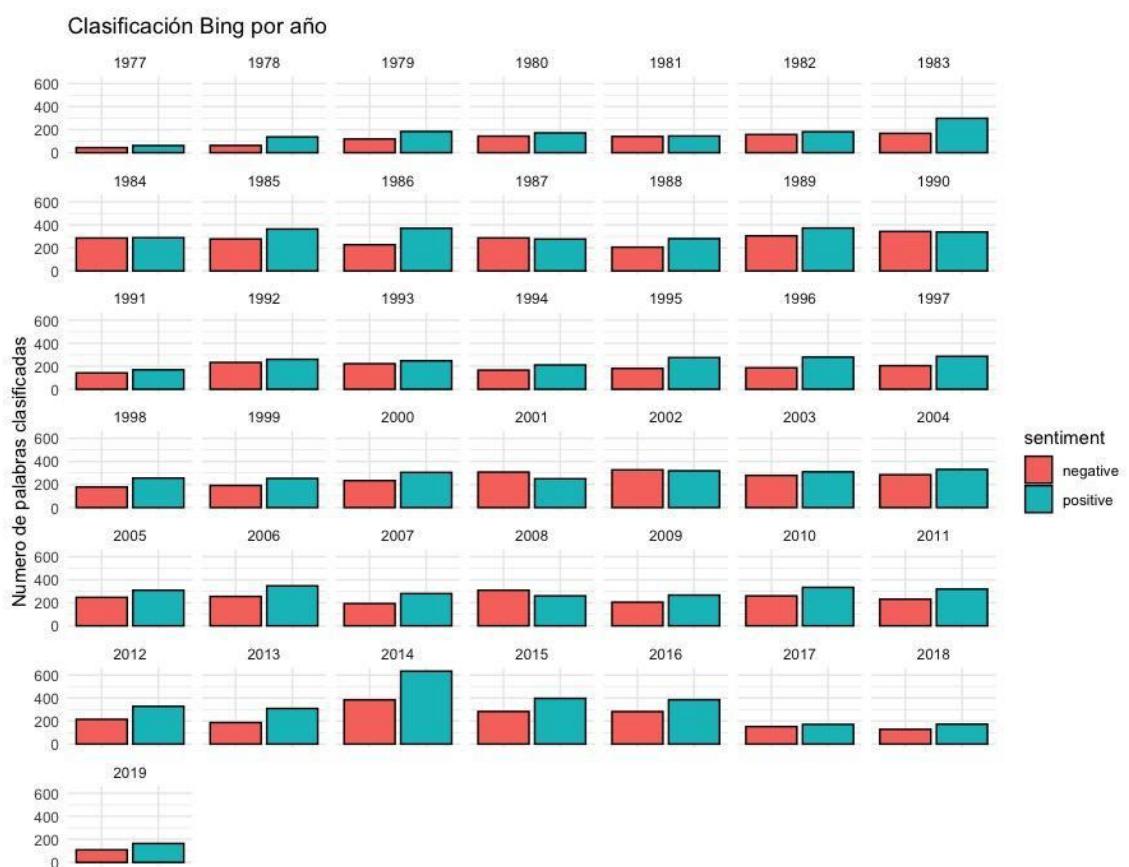


Figura 25: Distribución por años - Sentimiento Nrc (dividido por año)

4.5 Combinación de los 4 léxicos

Para concluir este capítulo, se procedió a comparar cómo actúan los 4 léxicos en una sola visualización. Para la obtención de esta figura se realizó el cálculo únicamente de palabras clasificadas como positivas restando las negativas (figura 26).

Comenzando con el léxico Affin, se manifiesta un sentimiento general positivo con los dos datos atípicos ya analizados en el capítulo 4.1. La carta perteneciente a 2001 que arroja un score negativo y la carta de 2008 con un score próximo al valor neutro.

Centrándose en Bing, se puede notar un sentimiento positivo también aunque observamos un aumento en el número de cartas negativas. En esta ocasión contamos con 5 cartas negativas y 2 positivas próximas al valor neutro.

Continuando con Loughran, se reconoce un cambio total de sentimiento, pasando de la positividad acostumbrada a la negatividad máxima. Si bien este resultado tiene una explicación interesante. En el punto 4.2 dónde se analizó este léxico, se comprobó como una vez aplicado a las cartas, el número de palabras negativas superaba al número de positivas (7194 vs 5235). El score obtenido para este sentimiento es claramente negativo en su mayoría, únicamente se manifiestan dos datos atípicos en 1977 y 1978.

Para concluir, se muestra como el léxico Nrc posee un sentimiento positivo en todos los años. Nótese que en esta clasificación, 9349 palabras eran negativas frente a 11902 positivas.

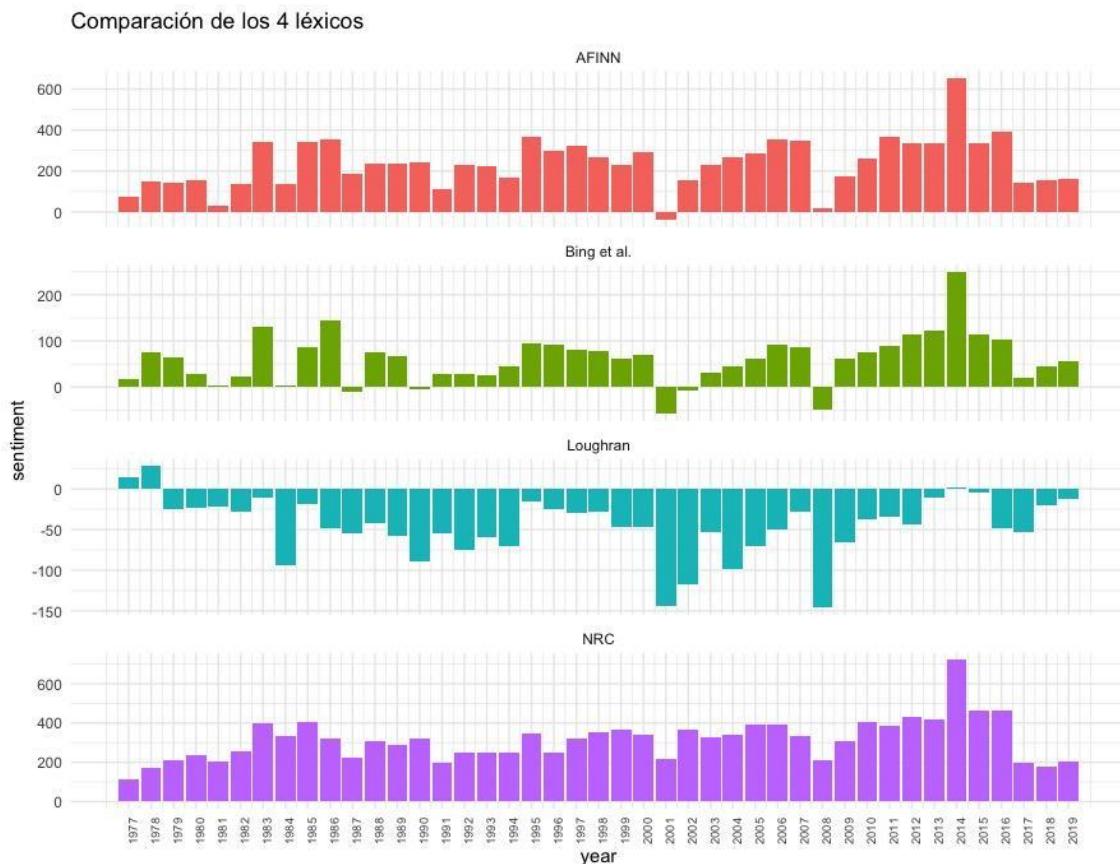


Figura 26: Aplicación de los 4 léxicos

5. RELACIONES ENTRE PALABRAS

En capítulos anteriores se centró la atención en el análisis de palabras tratadas de forma individual, sin tener en cuenta el conjunto. Existen otros métodos en los que se aplican técnicas para buscar relaciones entre palabras, ya sea por que se encuentran presentes en un mismo documento o por su propia consecución.

Se introduce el término de ‘ngrams’ haciendo referencia al análisis por conjuntos de ‘n’ palabras. Se hará uso de las librerías, ya mencionadas anteriormente, tidytext y tidyverse, así como también de ggplot2 (Pedersen 2017), para gráficos más avanzados de los que nos muestra ggplot2 y tidyverse para todo lo relacionado con correlaciones.

Se parte de la misma la función conocida, unnest_tokens(), pero esta vez configurando el apartado de token como se muestra a continuación:

```
letters %>%
unnest_tokens(bigram, text, token = "ngrams", n = 2)
```

Se establece n = 2 para examinar pares consecutivos o bigrams, consiguiendo así un modelo de relación entre la primera y la segunda palabra gracias a la frecuencia del propio conjunto.

Se presentan los 10 bigrams más frecuentes en el conjunto de cartas. Destaca observar la cantidad de stopwords presentes, por lo que se procede a su eliminación (tabla 15).

Tabla 15: Bigrams

	bigram	n
1	of the	2244
2	in the	2209
3	will be	919
4	of our	917
5	to the	905
6	we have	782
7	to be	770
8	and i	735
9	we will	688
10	for the	647

A continuación se dividió cada bigram en dos palabras independientes aplicando las funciones ungroup() y separate(). Se aplicó el filtro de stopwords consiguiendo una nueva lista de bigrams (tabla 16) y obteniendo así 56.168 bigrams

Tabla 16: Bigrams filtrados

	year	word1	word2	n
1	1983	tangible	assets	23
2	1994	scott	fetzer	22
3	1986	scott	fetzer	21
4	2007	pre	tax	19
5	1983	intrinsic	business	18
6	1983	economic	goodwill	16
7	2011	pre	tax	16
8	1983	blue	chip	15
9	2010	pre	tax	15
10	2012	net	worth	15

Se constata que los bigrams más comunes hacen referencia a ‘Scott Fetzer’ (empresa perteneciente al conglomerado de Berkshire), ‘asientos contables’, o ‘antes de impuestos’.

Se combinaron de nuevo las palabras con la función unite() para continuar trabajando con los bigrams.

5.1 Utilizar Bigrams para obtener sentimientos en contexto

Una de las grandes opciones que permiten los bigrams es obtener más contexto de los documentos que con las palabras individuales en sí. Aquella oración que se utilizó de ejemplo en el inicio del capítulo 5: “Mi hermano no es bueno”, el bigram “no bueno” será clasificado como negativo en esta ocasión.

Se aplicó el léxico Afinn (que permitía obtener un score) y se fijó uno de los dos términos a la palabra ‘not’. De esta forma se muestran los 20 bigrams que más contribuyen al sentimiento positivo y negativo (figura 27). Nótese la forma de interpretar este gráfico; se designa cómo de negativo es clasificado cada bigram, a mayor score más negativo y viceversa.

En el primer caso se presenta ‘not worth’ o ‘not good’ con un score alto indicando su alta contribución a la negatividad. Por otra parte se observa ‘not charged’ o ‘not worry’ con score menor que cero, por lo que se interpretan como positivos.

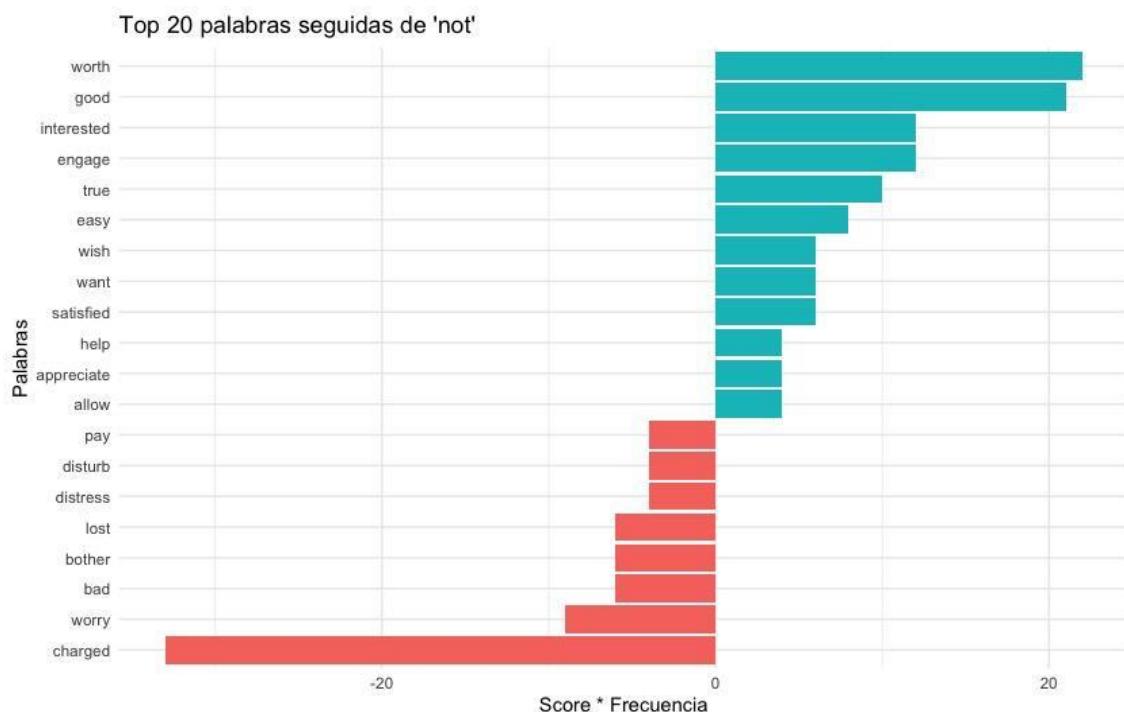


Figura 27: Top 20 palabras seguidas de ‘not’

Para continuar ampliando el análisis se incluyeron más términos negativos como ‘can’t’, ‘don’t’, ‘no’, ‘without’ y ‘won’t’. Nótese el resultado de las palabras que más contribuyen al sentimiento negativo, precedidas por una negación (figura 28).

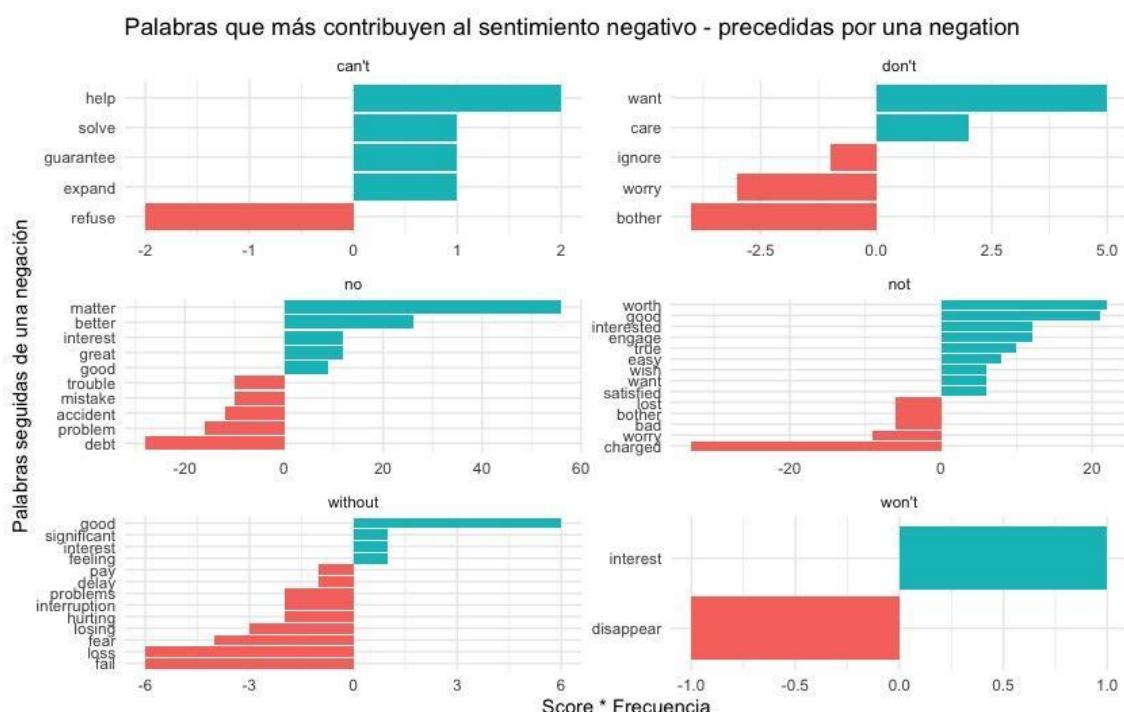


Figura 28: Palabras que más contribuyen al sentimiento negativo

5.2 Aplicando “graph” a nuestros bigrams

Se inicia otro apartado destinado a visualizar más allá de dos términos juntos. Se trata de graficar todas las relaciones a la vez. En esta ocasión se hará uso de la librería ‘igraph’, que proporciona una serie de funciones para la creación y control de grafos con facilidad.

La función utilizada fue ‘graph_from_data_frame()’. El primer paso consistió en filtrar a partir de una frecuencia, en este caso se fijó n = 40 (tabla 17).

Tabla 17: Bigrams aplicando graph

```
> bigram_graph <- bigram_counts %>%
+   filter(n > 40) %>%
+   graph_from_data_frame()
> bigram_graph
IGRAPH c834c68 DN-- 115 78 --
+ attr: name (v/c), n (e/n)
+ edges from c834c68 (vertex names):
 [1] pre      ->tax      net      ->worth
 [3] operating ->earnings berkshire ->hathaway
 [5] annual    ->meeting   balance   ->sheet
 [7] tax       ->earnings underwriting->profit
 [9] insurance ->business capital  ->gains
[11] annual    ->report    coca     ->cola
[13] scott     ->fetzer    insurance ->operations
[15] reported   ->earnings furniture ->mart
+ ... omitted several edges
```

A continuación se incorporó la librería ggraph, para adquirir una mejora de la propia visualización de igraph.

Para la construcción de un gráfico con ggraph se requieren varias capas, como mínimo una para los nodos, otra para los bordes y otra para el texto: (figura 29)

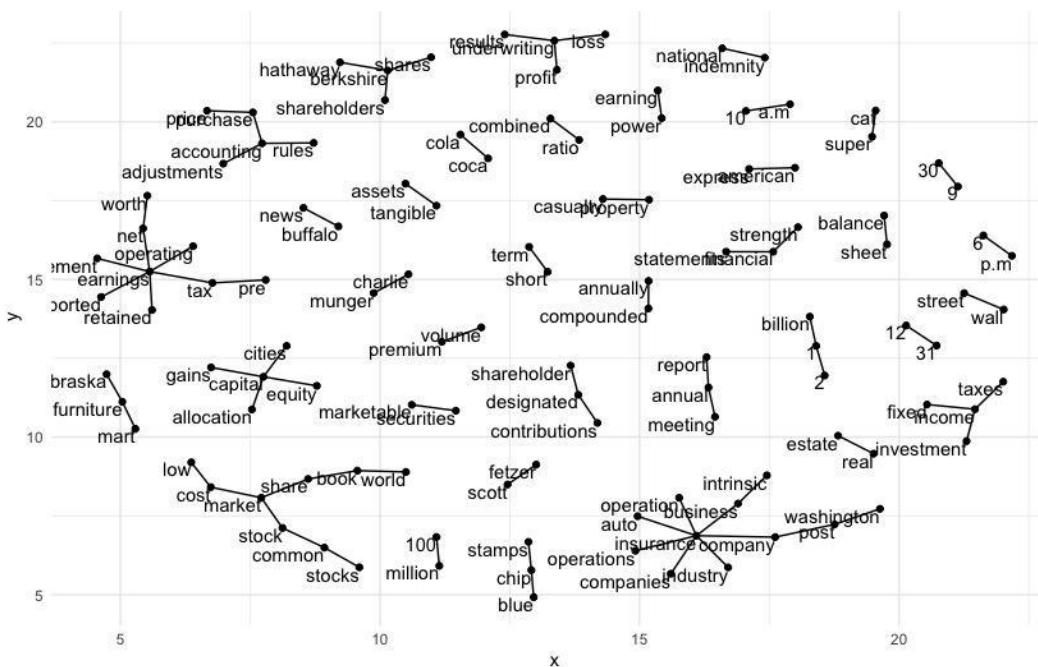


Figura 29: Aplicación de ggraph a nuestros bigrams.

Se aplicaron una serie de correcciones para mejorar la visualización (figura 30):

- Capa edge_alpha para hacer que los enlaces sean transparentes según lo común que sea el bigram
 - Agregación de flechas para conseguir una direccionalidad. Se construye usando grid :: arrow (), que incluye una opción end_cap que le indica a la flecha que finalice antes de tocar el nodo
 - Modificación de tamaño de puntos.
 - Asignación del tema theme_void ()

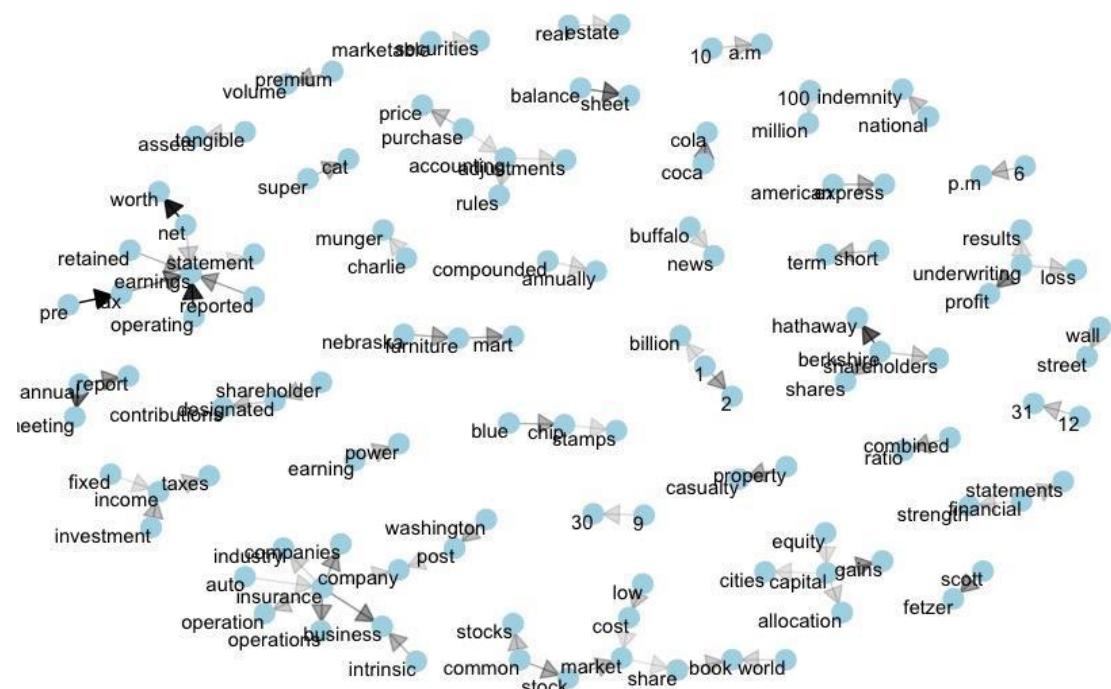


Figura 30: Aplicación de ggraph mejorado a nuestros bigrams

Se observa así cómo cada palabra depende de la anterior, por ejemplo, la palabra ‘bershkire’ puede estar seguida de ‘hathaway’, ‘shareholders’ o ‘shares’ en función de una mayor o menor frecuencia.

Nótese que se ha filtrado por frecuencia superior a 40 para una mejor visualización de los resultados.

Se presenta otro ejemplo de visualización (figura 31).

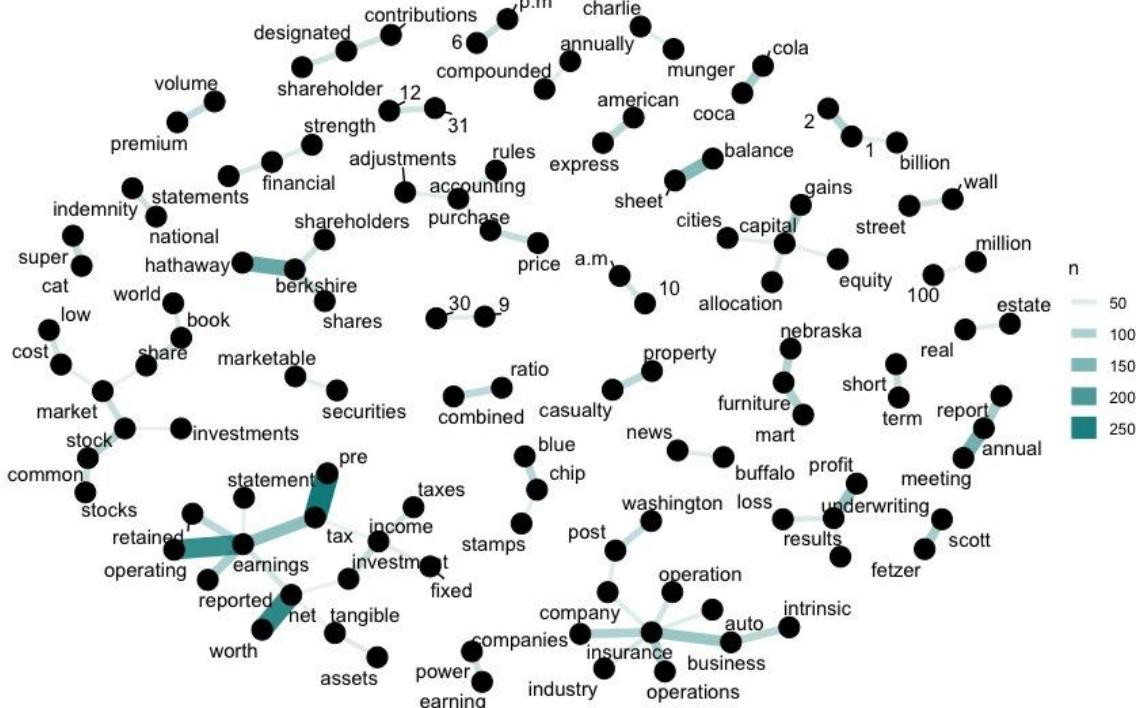


Figura 31: Aplicación de ggraph mejorado a nuestros bigrams (2)

5.3 Correlación de pares no consecutivos

En ocasiones se puede comprobar que tan correladas están dos palabras, independientemente de su posición en el documento. En este punto se tienen que tratar los datos de forma diferente.

El paquete `widyr` (David Robinson, 2016) facilita operaciones tales como cómputos y correlaciones (Figura 32). Se emplean un conjunto de funciones que hacen comparaciones por pares entre grupos de observaciones (por ejemplo, entre documentos o secciones de texto).

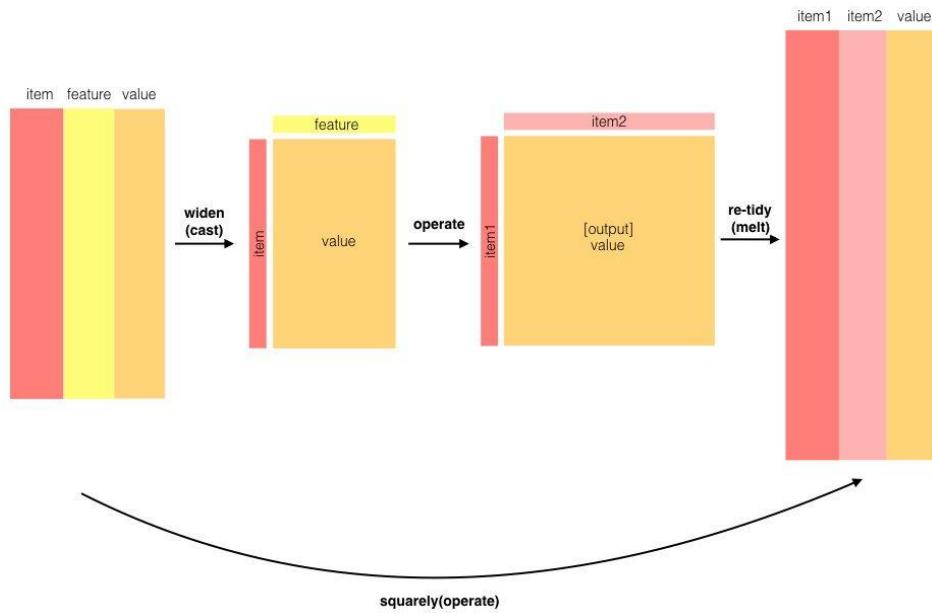


Figura 32: La filosofía detrás del paquete `widyr`

Fuente: Julia Silge and David Robinson (2020) “Text Mining with R - A Tidy Approach”

5.3.1 Recuento y correlación entre secciones

Una función útil de `widyr` es la función `pairwise_count()`. El prefijo `pairwise_` significa que dará como resultado una fila para cada par de palabras (ítem 1 y ítem 2). Esto permite contar pares de palabras comunes que aparecen conjuntamente en el mismo documento. (tabla 18)

Tabla 18: Correlación entre palabras

> <code>word_pairs</code>			
# A tibble: 75,066,454 x 3			
	item1	item2	
1	berkshire	letter	43
2	hathaway	letter	43
3	operating	letter	43
4	earnings	letter	43
5	share	letter	43
6	capital	letter	43
7	gains	letter	43
8	company	letter	43
9	losses	letter	43
10	insurance	letter	43
# ... with 75,066,444 more rows			

Este también es un formato ordenado, pero de una estructura muy diferente que puede ser usado para responder a nuevas cuestiones.

5.3.2 Correlación Pairwise

Pares como ‘berkshire’ y ‘letter’ son las palabras más comunes, pero eso no es particularmente significativo. En su lugar, se puede querer examinar la correlación entre las palabras, lo que indica la frecuencia con la que aparecen juntas en relación con la frecuencia con la que aparecen por separado.

En particular, aquí hizo uso del coeficiente phi, una medida común para la correlación binaria (tabla 19).

Tabla 19: Correlación entre palabras

	Has word Y	No word Y	Total
Has word X	n_{11}	n_{10}	$n_{1\cdot}$
No word X	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	n

n_{11} representa el número de documentos donde la palabra X y la palabra Y aparecen juntas, n_{00} los documentos donde ninguno aparece. n_{10} y n_{01} documentos donde aparece una si y la otra no. Podríamos definir el coeficiente phi (equivalente a Pearson) de la siguiente manera:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 0}n_{\cdot 1}}}$$

Figura 33: Coeficiente de phi

La función pairwise_cor() en widyr permite encontrar el coeficiente phi entre palabras en función de la frecuencia con la que aparecen en la misma sección. Su sintaxis es similar a pairwise_count().

Este formato de salida es útil para la exploración. Por ejemplo, se procedió a encontrar las palabras más correlacionadas con la palabra "pounds" usando un filter() (tabla 20).

Tabla 20: Correlación entre palabras

item1	item2	correlation
1 pounds	store	0.563
2 pounds	pairs	0.542
3 pounds	minutes	0.491
4 pounds	materially	0.491
5 pounds	answer	0.489
6 pounds	furniture	0.489
7 pounds	mart	0.489
8 pounds	hour	0.489
9 pounds	bob	0.471
10 pounds	repeat	0.463
# ... with 1,935 more rows		

Probamos ahora con ‘customer’ y con ‘fear’(figura 34).

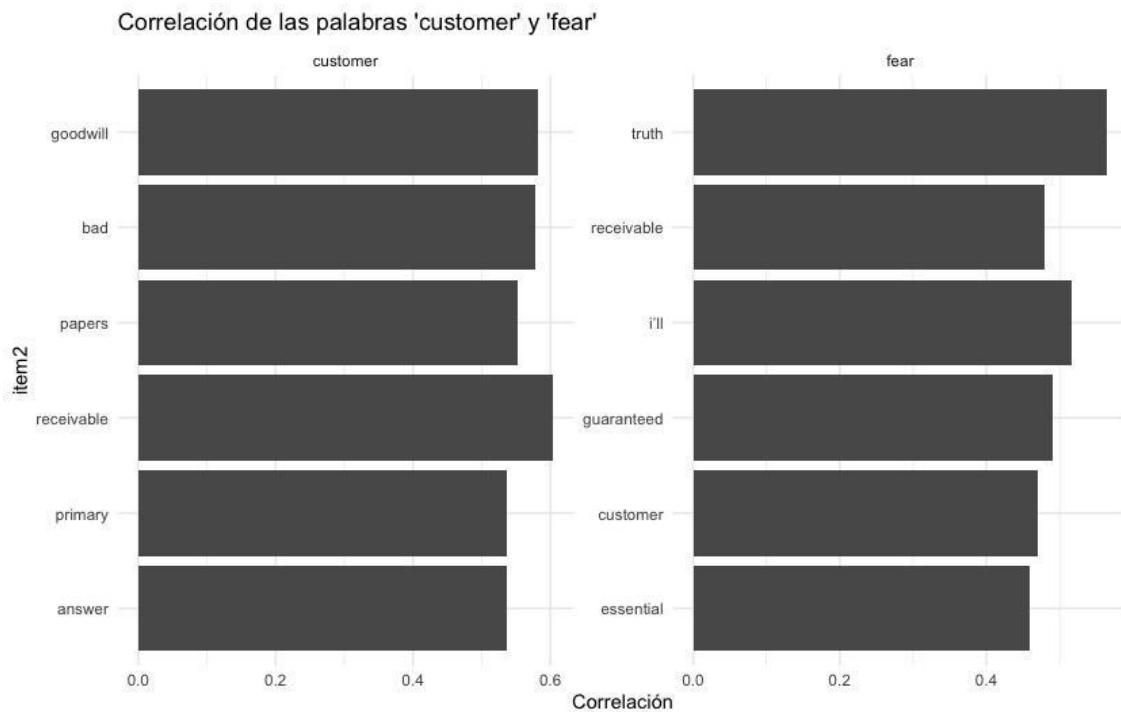


Figura 34: Mayores correlaciones de ‘customer’ y ‘fear’

Se ha mostrado cómo el enfoque de texto ordenado es útil no sólo para analizar palabras individuales, sino también para explorar las relaciones y conexiones entre palabras. Dichas relaciones pueden involucrar n-gramas, lo que permite ver qué palabras tienden a aparecer después de otras, o coincidencias y correlaciones, para palabras que aparecen cerca unas de otras.

5.4 Trigrams

Continuando con las relaciones entre palabras, se pueden generar una agrupación de 3 términos consecutivos o trigrams. Se muestra un top 10 por frecuencia (tabla 21) acompañado de la visualización (figura 35).

Tabla 21: Top 10 trigrams

> letters_trigrams	word1	word2	word3	n
1	pre	tax	earnings	110
2	nebraska	furniture	mart	91
3	shareholder	designated	contributions	65
4	washington	post	company	49
5	blue	chip	stamps	45
6	coca	cola	company	39
7	net	investment	income	39
8	million	pre	tax	38
9	common	stock	investments	37
10	net	tangible	assets	37

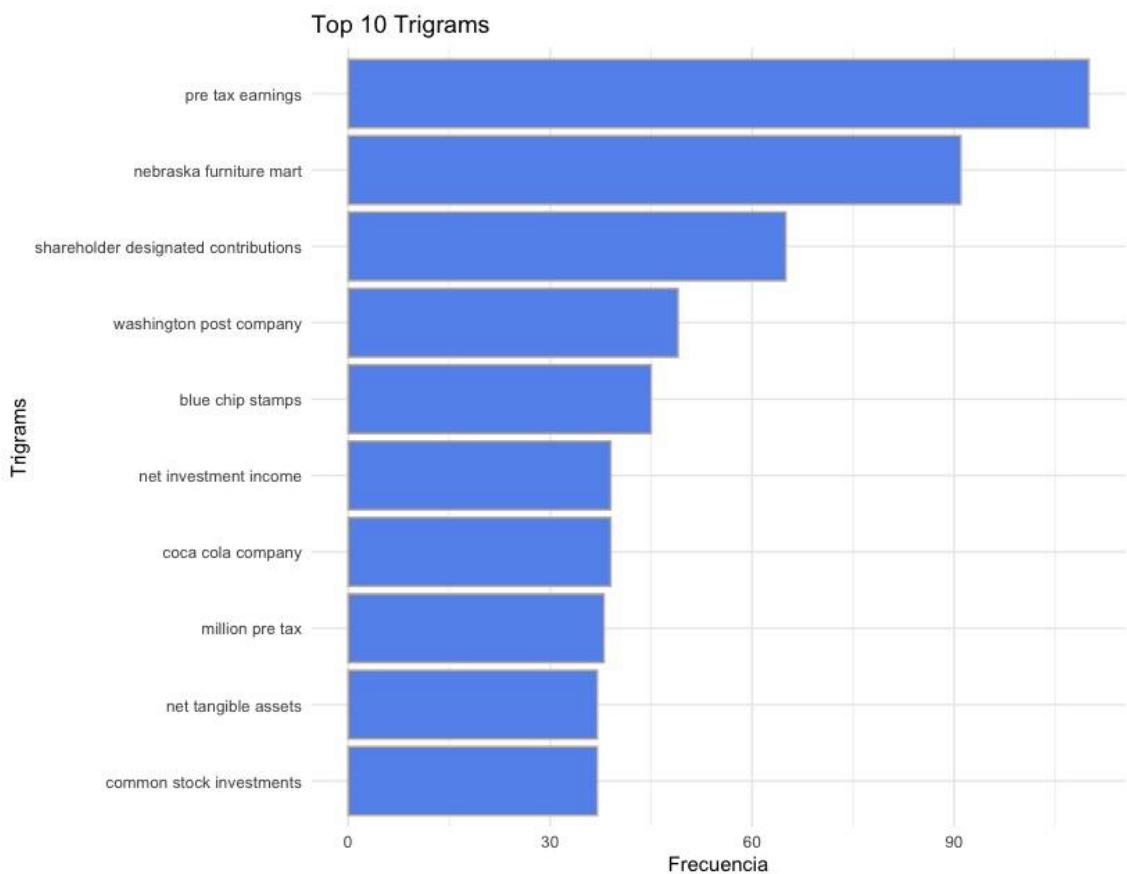


Figura 35: Mayores correlaciones de ‘customer’ y ‘fear’

6. INVERSE DOCUMENT FREQUENCY

La pregunta central en la minería de texto y el procesamiento del lenguaje natural es cómo cuantificar de qué trata un documento. En anteriores capítulos se experimentó con la frecuencia de término. A pesar de que normalmente aplicamos el filtro de ‘stopwords’ para su eliminación, es posible encontrar que alguna de estas palabras tengan una importancia diferente según el documento a analizar.

Se presenta un nuevo enfoque, la frecuencia de documentos inversa (idf) de un término. Su funcionamiento consiste en aplicar un peso (importancia) mayor a las palabras poco comunes y disminuirlo en las más frecuentes. Si se combina esta técnica con la frecuencia de término aparece el tf-idf de un término (multiplicación de ambas), la frecuencia de un término ajustada por la poca frecuencia con la que se usa.

Se define el tf-idf como:

$$idf(\text{term}) = \ln \left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}} \right)$$

Figura 36: Expresión matemática ‘tf-idf’

6.1 Frecuencia de término en las cartas del señor Warren buffet

Se comenzó mirando de nuevo el conjunto de cartas y examinando la frecuencia de cada término. Se crea una nueva variable ‘each_letter_words’ y se aplican funciones del paquete dplyr como group_by() y join(). Se añade también la columna ‘total’ que hace referencia al total por cada año.

Tabla 22: Frecuencia de términos

	year	word	n	total
1	2014	the	914	23681
2	1990	the	760	15350
3	2015	the	699	16918
4	1986	the	686	13644
5	2002	the	685	14551
6	1989	the	668	14417
7	2016	the	651	16099
8	2004	the	646	14614
9	1985	the	643	13751
10	2012	the	622	14146

En cada fila se presenta una única palabra para cada carta en su respectivo año; de esta forma se obtiene ‘n’ (frecuencia individual en ese año) y total (frecuencia total de palabras en cada carta). Se observa como la palabra ‘the’ es la más repetida ocupando las 10 primeras posiciones.

El tf es calculado como n/total para cada carta, el número de veces que aparece una palabra dividido por el número total de términos (figura 37). Se observa que para el conjunto total, existen términos muy frecuentes y otros difícilmente pueden ser cuantificados. Se divide con facets() por cada año observando una repetición prácticamente igual en cada carta (figura 38)

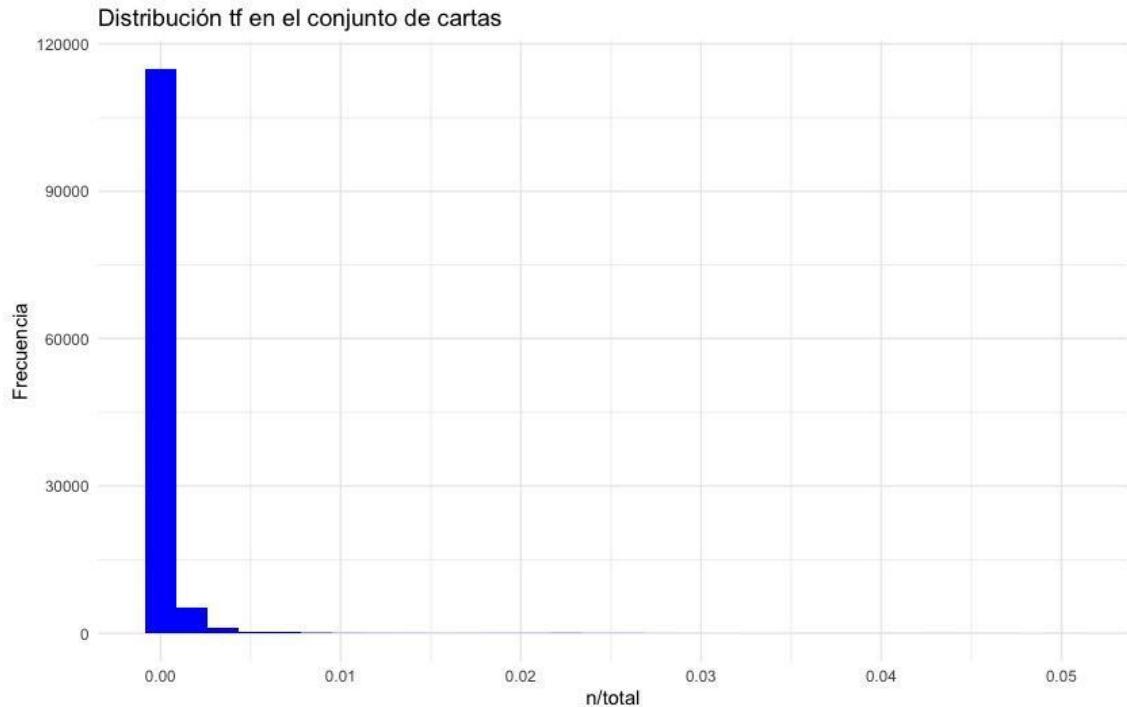


Figura 37: Frecuencia de términos en las cartas

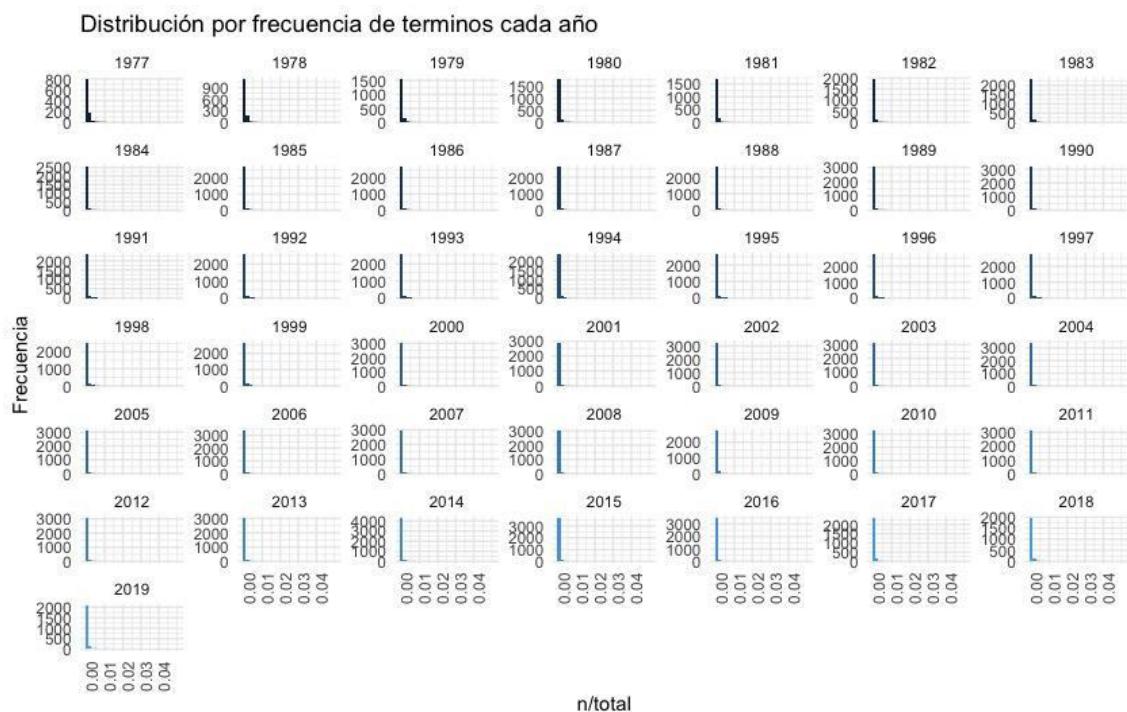


Figura 38: Frecuencia de términos en cada carta.

6.2 Ley de Zipf

Las distribuciones mostradas en el punto anterior son las más comunes en el lenguaje natural. Se introduce ahora la “ley de Zipf”, en honor a George Zipf, lingüista estadounidense (siglo XX).

En la mayoría de los textos escritos la palabra más frecuente suelen ser el doble de veces más utilizada que la segunda más utilizada y el triple de veces que la tercera.

Se examina la ley de Zipf para las cartas del señor Buffett (tabla 23).

Tabla 23: Aplicación de Zipf

	year	word	n	total	rank	term frequency`
	<int>	<chr>	<int>	<int>	<int>	<dbl>
1	2014	the	914	23681	1	0.0386
2	1990	the	760	15350	1	0.0495
3	2015	the	699	16918	1	0.0413
4	1986	the	686	13644	1	0.0503
5	2002	the	685	14551	1	0.0471
6	1989	the	668	14417	1	0.0463
7	2016	the	651	16099	1	0.0404
8	2004	the	646	14614	1	0.0442
9	1985	the	643	13751	1	0.0468
10	2012	the	622	14146	1	0.0440
	# ... with 122,313 more rows					

Ahora se tienen dos columnas nuevas: rank y term frequency. Rank hace referencia al rango otorgado a cada palabra en función de la frecuencia, un menor rango equivale a una alta frecuencia y baja importancia. Se visualizan estos resultados, eje x para rango, eje y para la frecuencia de términos en escala logarítmica. Destacar la pendiente negativa por la relación inversamente proporcional.

Ley de Zipf aplicado a nuestras cartas

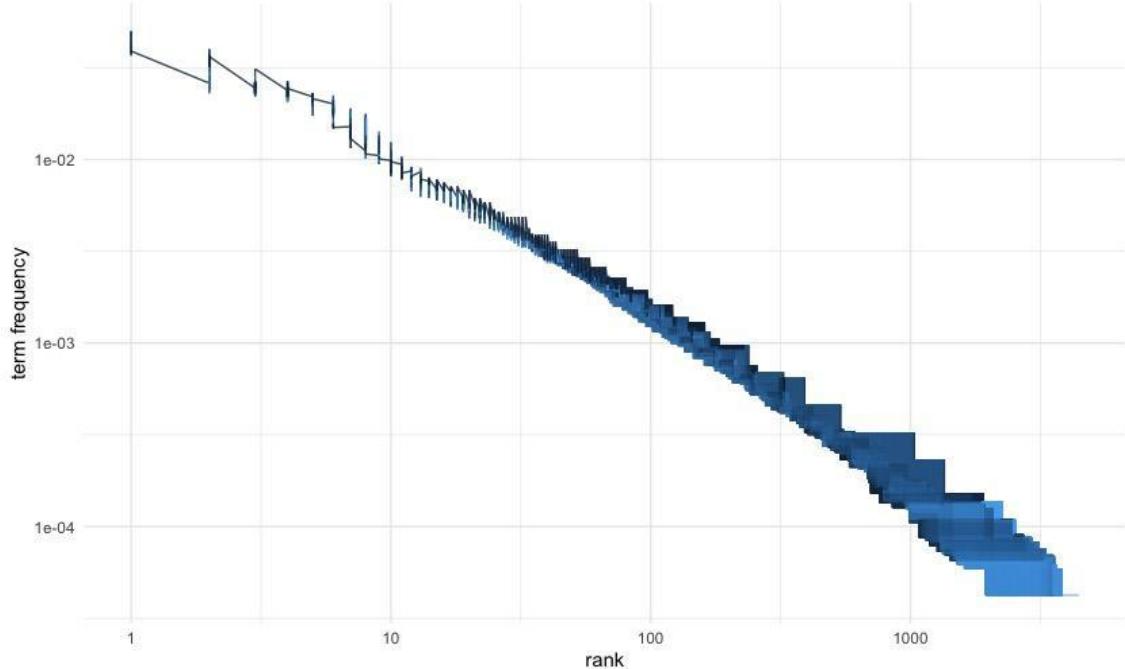


Figura 39: Visualización ley de Zipf

Ley de Zipf, división por años

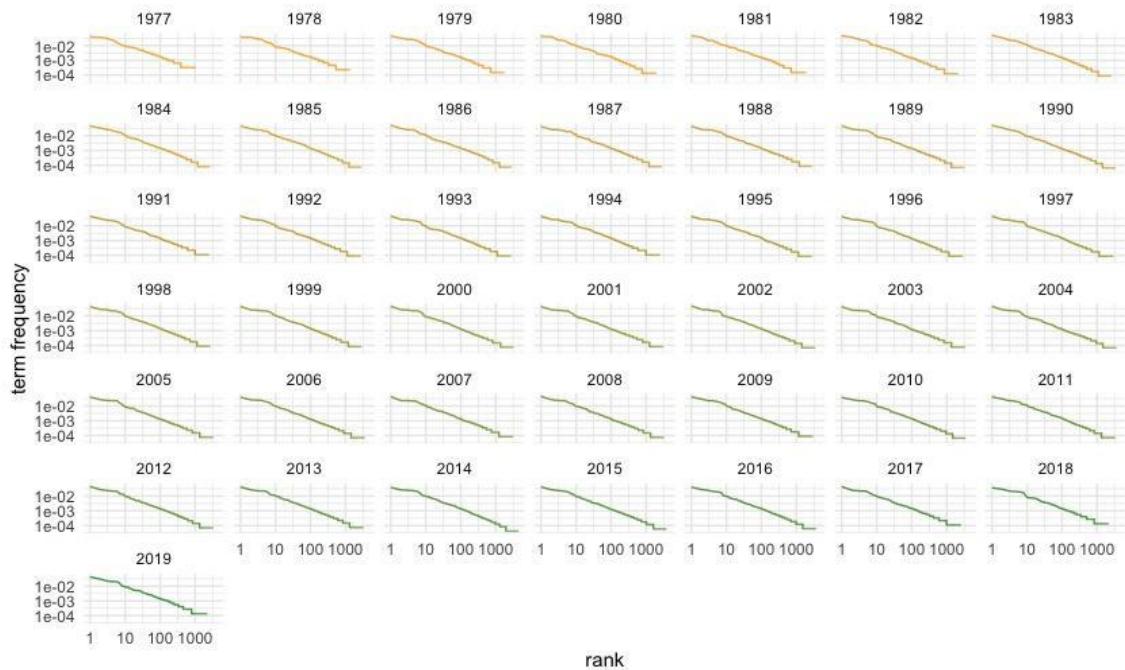


Figura 40: Visualización ley de Zipf por carta

Se observa que las 43 cartas son similares, se procede a aplicar la pendiente conjunta al gráfico (tabla 24 y figura 41).

Tabla 24: Exponente y pendiente

```
Call:
lm(formula = log10(`term frequency`) ~ log10(rank), data = rank_subset)

Coefficients:
(Intercept)  log10(rank)
-1.0544      -0.9037
```

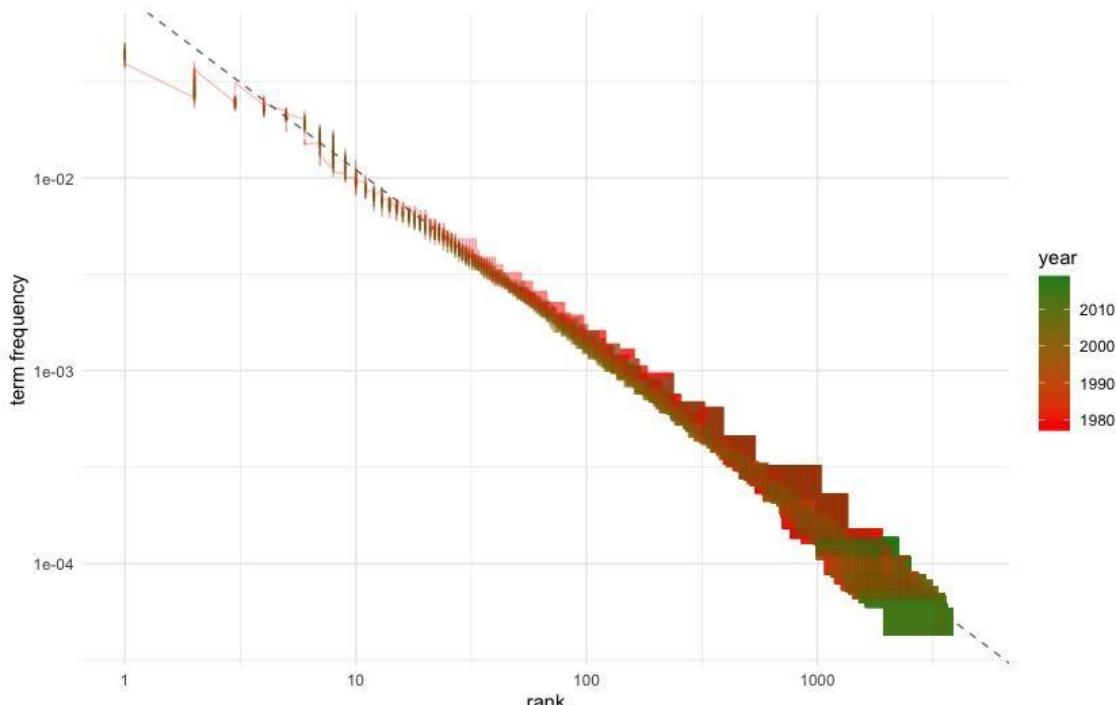


Figura 41: Visualización ley de Zipf Ajuste con un exponente

Se concluye este apartado destacando la homogeneidad en la distribución de términos en las 43 cartas. En un lenguaje natural es común encontrarnos esta situación, pocas palabras son repetidas en muchas ocasiones y muchas palabras se repiten muy pocas.

6.3 La función bind_tf_idf

Se ha visto como tf-idf busca encontrar las palabras importantes para el contenido de cada documento disminuyendo el peso de las palabras de uso común y aumentando el peso de las palabras que no se usan mucho en una colección. Calculando el tf-idf se intentan encontrar las palabras que son importantes en un texto, pero que no son demasiado comunes.

Con ayuda de la función bind_tf_idf del paquete tidytext se obtuvieron las columnas que interesan: tf, idf y tf_idf. Esta última fue ordenada de mayor a menor para observar las palabras más interesantes. En los casos de palabras comunes idf y tf_idf tienden a 0. Este enfoque disminuye el peso de las palabras comunes (tabla 25).

Tabla 25: tf_idf

	year	word	n	tf	idf	tf_idf
1	2017		2017	31	0.0033798517	2.1517622 0.007272637
2	2019		2019	19	0.0025970476	2.3749058 0.006167743
3	1978	safeco	9	0.0020794824	2.3749058 0.004938575	
4	1988	arbitrage	35	0.0029595806	1.5639755 0.004628712	
5	2018		2018	17	0.0022895623	1.9694406 0.004509157
6	2011		2011	42	0.0029799915	1.4586150 0.004346660
7	1986	nhp	15	0.0010993843	3.7612001 0.004135005	
8	2012		2012	39	0.0027569631	1.4586150 0.004021348
9	2016		2016	30	0.0018634698	2.1517622 0.004009744
10	1998	re's	14	0.0012288247	3.0680529 0.003770099	

Se procede a visualizar estos términos en el conjunto de cartas dividido por años (figura 42) y se hace zoom en algunos datos (figura 43).

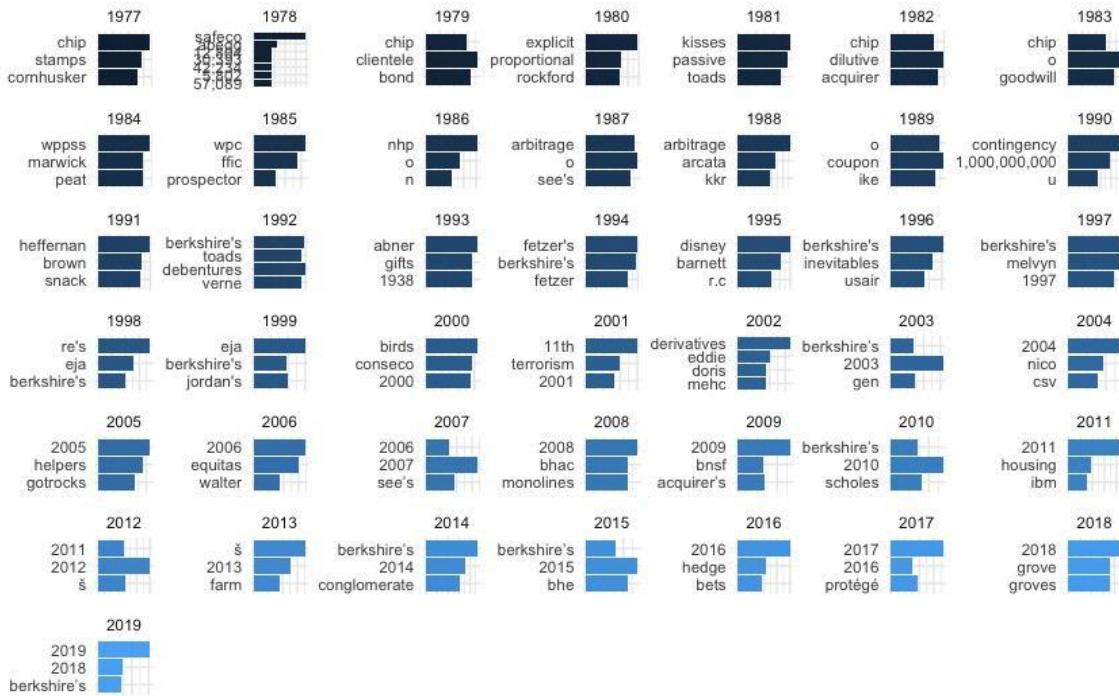


Figura 42: Visualización de tf_idf por año

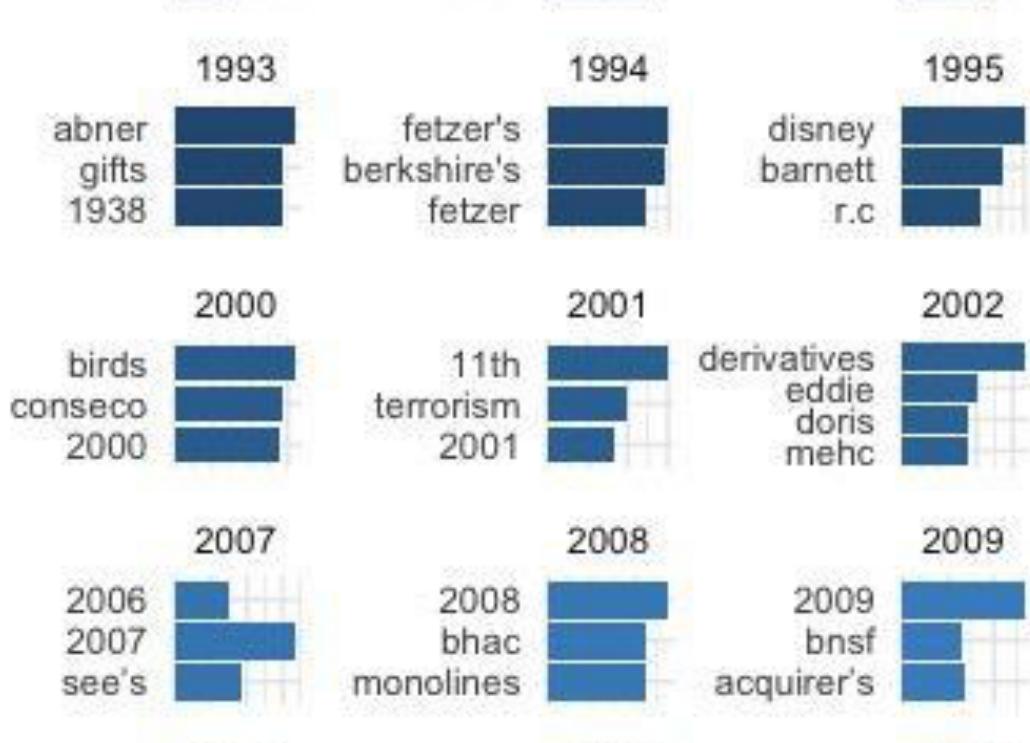


Figura 43: Visualización de tf_idf por año (zoom)

Se obtienen así las palabras que el estadístico considera ‘más relevantes’ para cada año. Se observa por ejemplo, en el año 2001, se da mucha importancia a los ataques terroristas del 11 de Septiembre. Destaca también como importante el propio año que se menciona en cada carta, entendiendo que es un término importante ya que se hace referencia a él en su contenido financiero.

Se eliminaron las stopwords y algún término correspondiente al año en concreto para observar nuevos resultados (figura 44). Destaca la relevancia que se le da en los años 2008 y 2009 a los productos derivados, muy probablemente fruto de la crisis vivida en ese período (figura 45).



Figura 44: Visualización de tf_idf por año (sin stopwords)

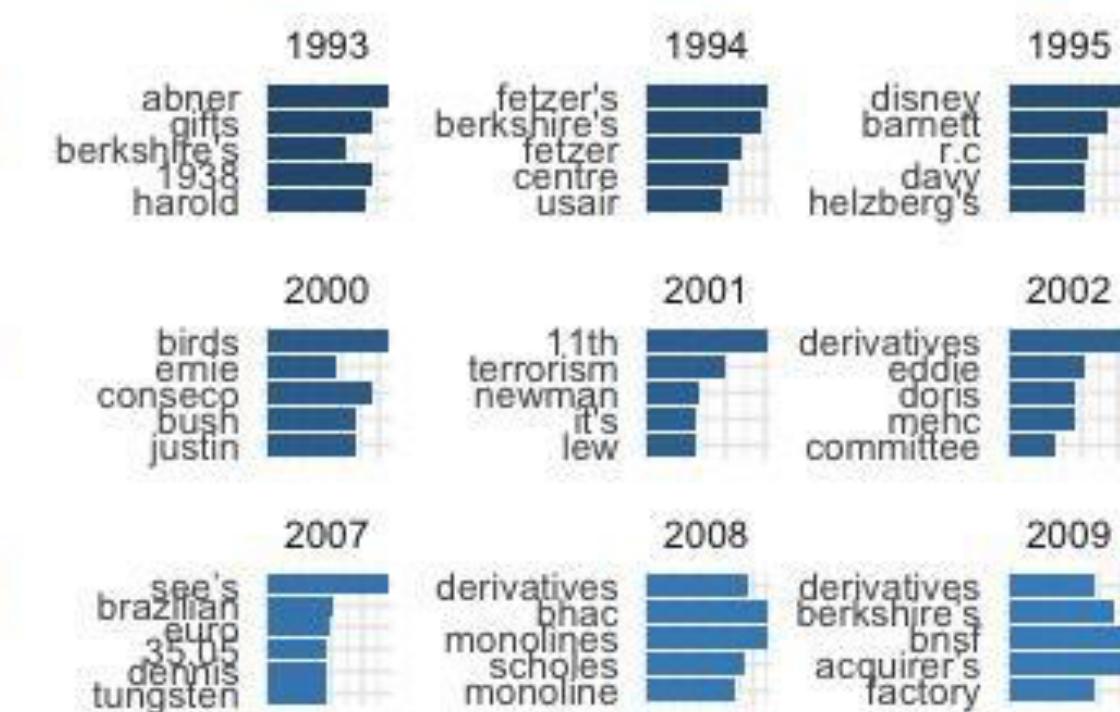


Figura 45: Visualización de tf_idf por año (sin stopwords y zoom)

7. TOPIC MODELLING

En nuestro día a día es frecuente encontrarnos con conjuntos extensos de documentos sin una precisa clasificación. Se puede tener interés en aplicar unos patrones para así dividir el total en una serie de temas específicos y facilitar su comprensión. Aparece entonces el Topic Modelling como técnica automática no supervisada cuya función es la de agrupación natural de los elementos y, en ocasiones, sin conocer de antemano cuales pueden ser esos temas.

Se introduce el concepto de LDA (latent dirichlet allocation), un modelo generativo probabilístico muy conocido para este fin. Es muy útil cuando la colección de documentos es muy extensa. Consiste en un modelo bayesiano jerárquico a tres niveles, en el que cada elemento de la colección de textos es modelado como una mezcla finita sobre un conjunto subyacente de temas. En cambio, cada tema es modelado como una mezcla infinita sobre un conjunto subyacente de probabilidades.

- Cada documento consiste en una mezcla de temas. Por ejemplo, "La carta del año 1980 contiene el 90% del tema X y el 10% del tema Y, mientras que la carta de 1990 contiene el 30% del tema X y el 70% del tema Y".
- Cada tema es una mezcla de palabras. Por ejemplo, podríamos imaginar el tema de "economía" y "medicina". Las palabras más comunes en el tema económico podrían ser "Beneficios", "Accionistas" o "Inflación", mientras que el tema médico pueden aparecer palabras como "Salud", "Hospital", "Enfermedades". En ocasiones tendremos palabras que pertenecen a ambos, por ejemplo "Sociedad".

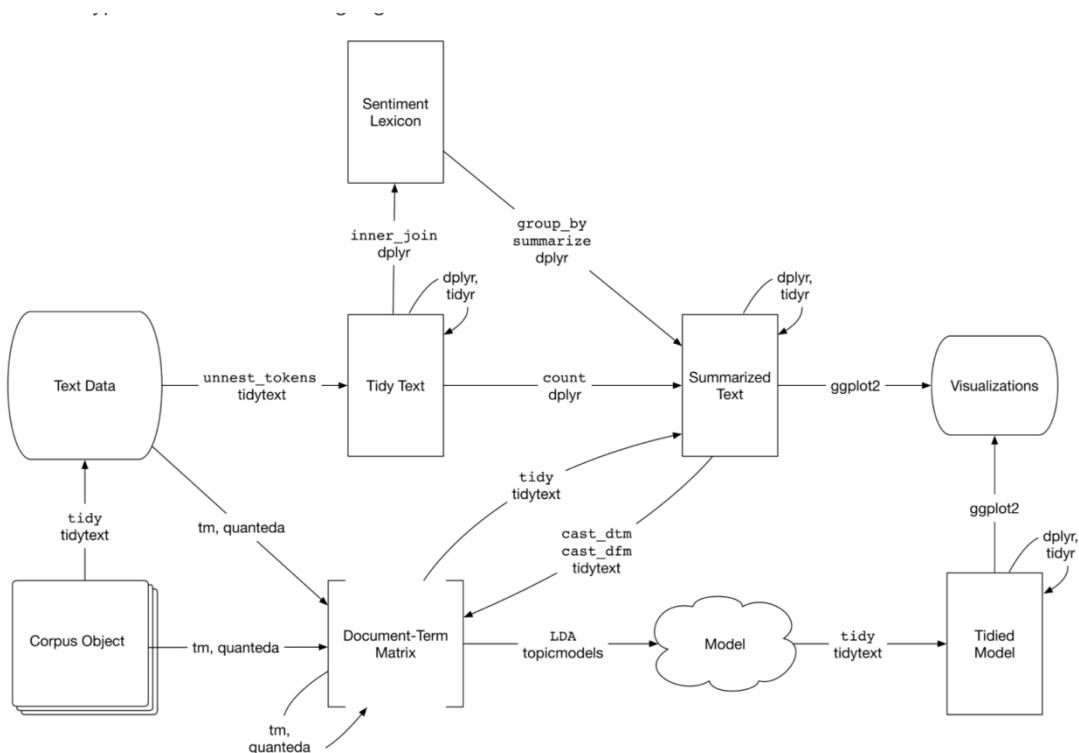


Figura 46: Diagrama de flujo de un análisis de texto

Fuente: Julia Silge and David Robinson (2020) "Text Mining with R - A Tidy Approach"

Se trata cada documento como una mezcla de temas y cada tema como una mezcla de palabras. Con esto se consigue una aproximación al lenguaje natural por medio de la superposición de los temas en cada documento. Para su correcta aplicación en R se hizo uso de la librería `topicmodels`, consiguiendo unos documentos ordenados posteriormente aplicables en `ggplot2` y `dplyr`.

Se parte del supuesto de que cada documento, en este caso cartas, consiste en una mezcla de temas, aunque a simple vista no sea fácil identificarlos. Se considera a esto como una estructura latente que se desea conocer. Este método ajusta una importancia relativa para cada tema.

Para la correcta aplicación e interpretación de LDA, debemos destacar que los diferentes temas están distribuidos por todos los documentos. Asignamos un tema a un documento por probabilidad.

7.1 LDA en años

Se crea la variable ‘letter_count’ para obtener un contador de frecuencia de cada palabra junto al año en el que aparece (tabla 26).

Tabla 26: Contador por año

	year	word	n
1	2014	berkshire	203
2	1985	business	112
3	1983	business	97
4	1984	business	96
5	2014	business	92
6	1990	business	90
7	2015	berkshire	90
8	1980	earnings	87
9	2016	berkshire	86
10	1989	business	85

En este momento los datos se encuentran en una forma ordenada (término por documento por fila), pero para la correcta aplicación del paquete `topicmodels` se requiere de la creación de la “DocumentTermMatrix”. Se convierte así una tabla de un token por fila en un DocumentTermMatrix con la función en `tidytext` de “`cast_dtm()`”. Se obtuvo la nueva variable ‘years_dtm’ (tabla 27). Se comprueba así como tenemos los 43 documentos.

Tabla 27: Creación de la matriz

```
> years_dtm
<<DocumentTermMatrix (documents: 43, terms: 15076)>>
Non-/sparse entries: 88397/559871
Sparsity           : 86%
Maximal term length: 25
Weighting          : term frequency (tf)
```

Se procedió a aplicar la función LDA() para crear un modelo de X temas (k = 16) (tabla 28). Después de numerosas pruebas con diferentes valores de ‘k’, aplicando filtros por frecuencia e incluso eliminando palabras muy repetidas, se ha llegado a la conclusión que el valor 16 es el que mejor resultados arroja para este ejercicio.

Tabla 28: LDA

```
> years_lda  
A LDA_VEM topic model with 16 topics.
```

Esta función devuelve un objeto que contiene los detalles completos del ajuste del modelo, cómo se asocian las palabras con los temas y cómo se asocian los temas con los documentos.

Continuar con el análisis implicó explorar e interpretar el modelo utilizando las funciones de ordenación del paquete tidytext. Se examinaron las probabilidades por tema por palabra.

El paquete tidytext proporciona el método para extraer las probabilidades por tema por palabra, denominado β ("Beta"), del modelo (tabla 29).

Tabla 29: Betas del modelo

```
> years_topics  
# A tibble: 241,216 × 3  
  topic term      beta  
  <int> <chr>    <dbl>  
1     1 berkshire 0.0101  
2     2 berkshire 0.00992  
3     3 berkshire 0.0164  
4     4 berkshire 0.00764  
5     5 berkshire 0.00744  
6     6 berkshire 0.0114  
7     7 berkshire 0.00580  
8     8 berkshire 0.00675  
9     9 berkshire 0.0124  
10    10 berkshire 0.0135  
# ... with 241,206 more rows
```

La probabilidad de que el término ‘berkshire’ sea generado en el topic 1 es de 0.0101 y en el 2 de 0.00992. Se hizo uso de top_n () de dplyr para encontrar los 15 términos principales dentro de cada tema. Al tratarse de datos ordenados, se procedió a visualizar con ggplot2 (Figura 47 y 48).

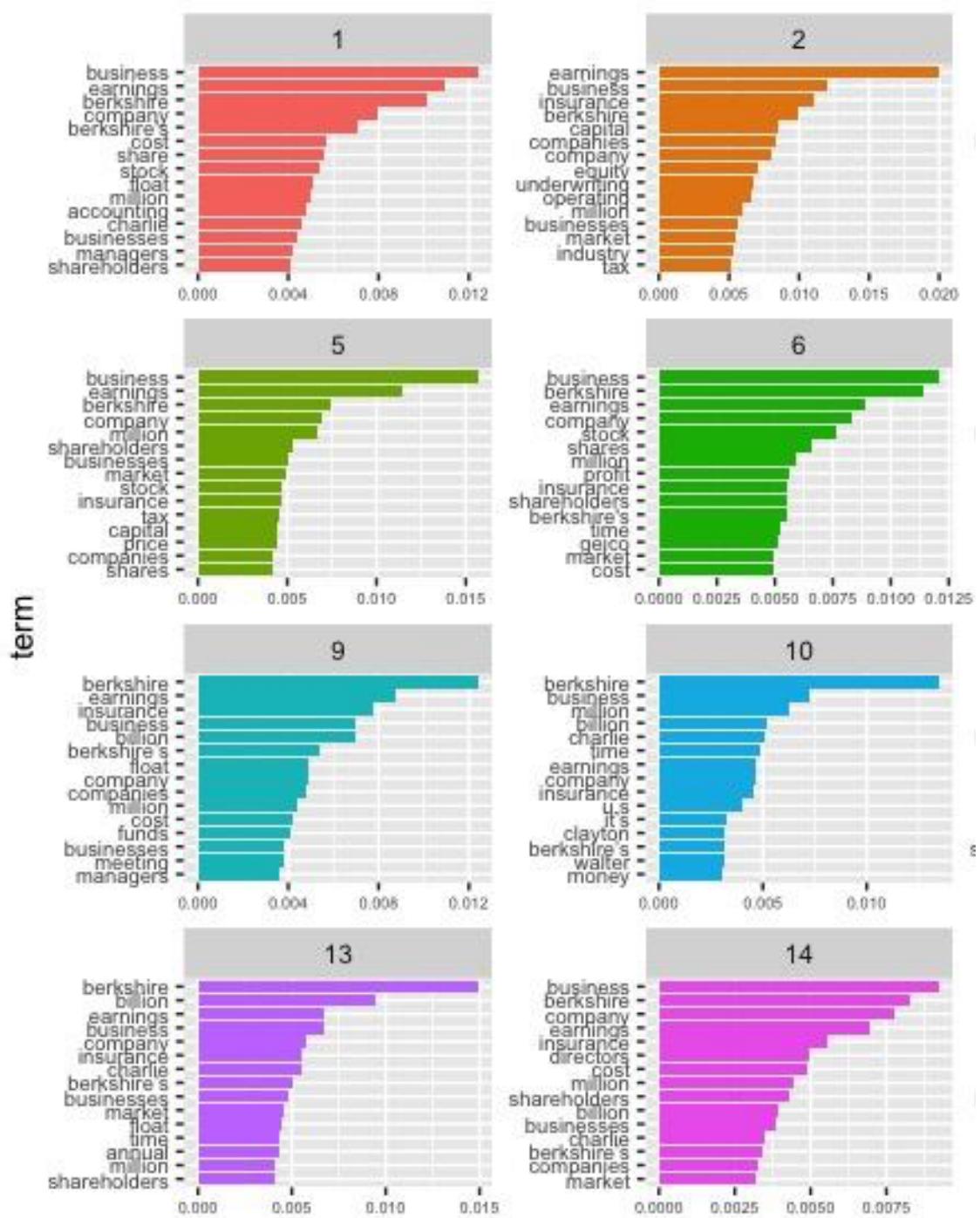


Figura 47: Clasificación en 16 temas (parte 1)

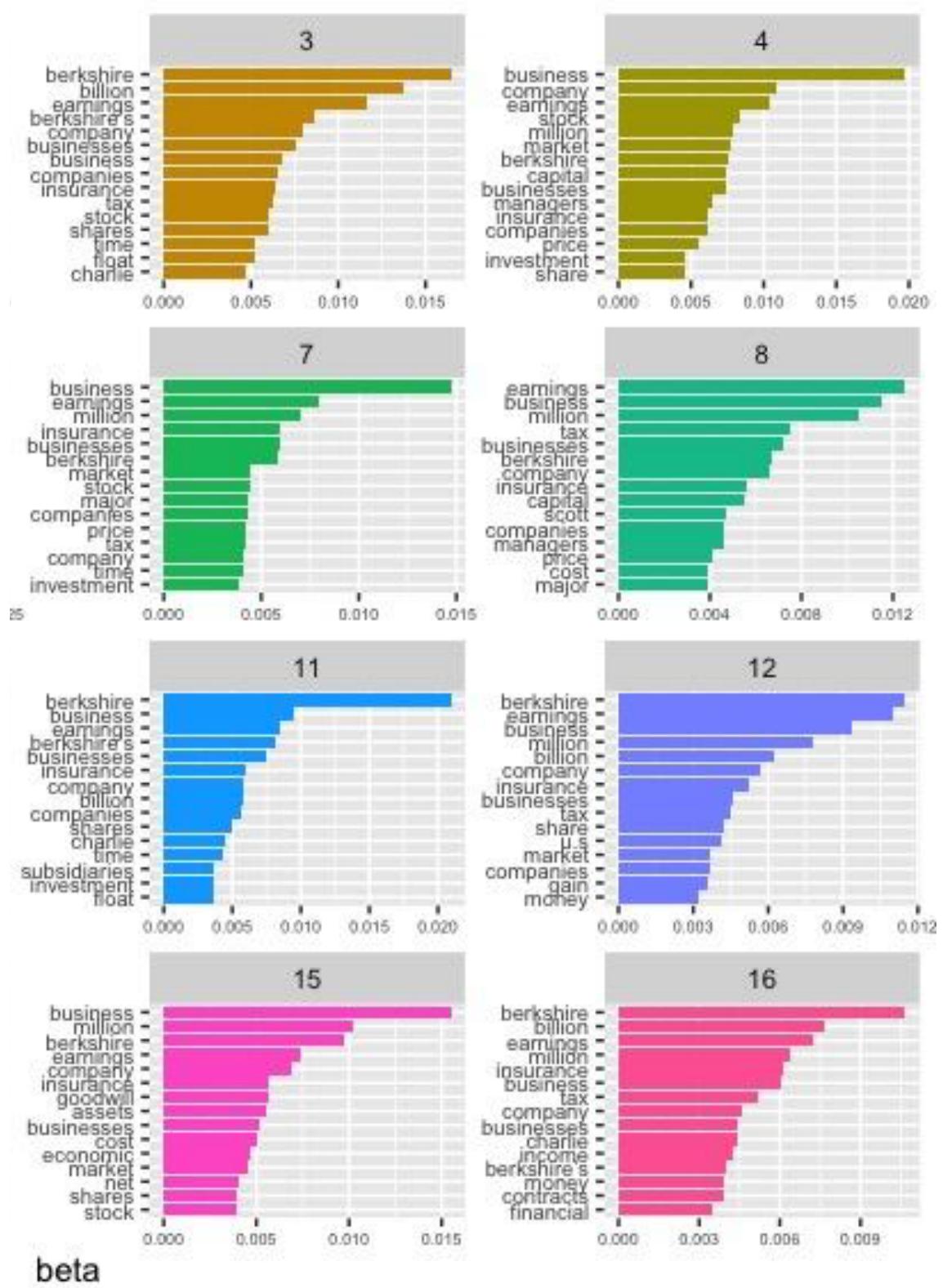


Figura 48: Clasificación en 16 temas (parte 2)

7.1.1 Ejemplo con dos temas

A modo de curiosidad y, a pesar de que se ha comprobado que con dos únicos temas el análisis no resulta eficiente, si que es interesante representar visualmente cuáles son las palabras (suponiendo $k = 2$) que más contribuyen a cada tema (figura 49).

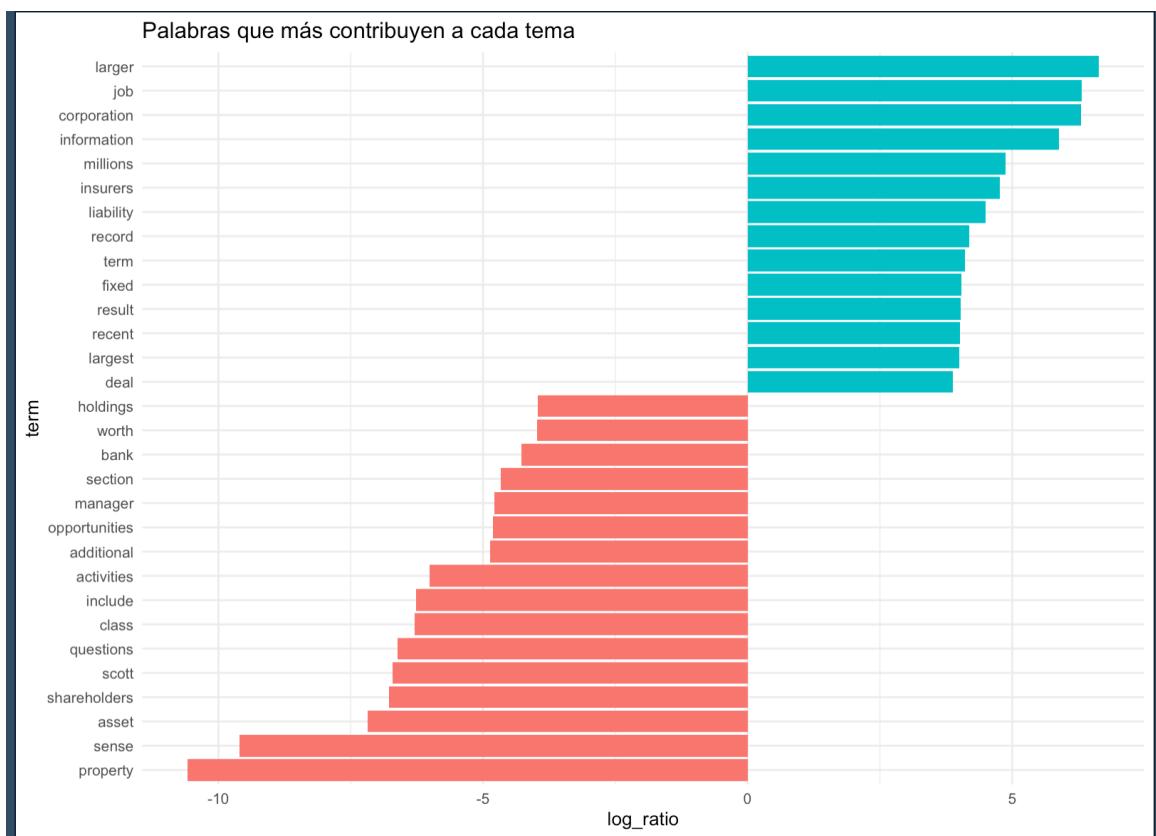


Figura 49: Clasificación en 2 temas

Aunque a simple vista pueda parecer muy similar, se puede destacar como el topic 1 (color rojo) tiende más a términos financieros, mientras que el topic 2 (color azul) tiende a los puramente económicos.

7.2 Probabilidades de pertenencia a un tema

Se retoman los 16 temas y se realiza la operación inversa, modelar cada documento como una mezcla de los temas existentes. Se examinaron las probabilidades por documento por tema, llamadas γ ("Gamma"), con el argumento matriz = "gamma". (tabla 30)

Tabla 30: Betas del modelo

	document	topic	gamma
	<chr>	<int>	<dbl>
1	2014	1	0.000000944
2	1985	1	0.00000166
3	1983	1	0.00000196
4	1984	1	0.00000181
5	1990	1	0.00000148
6	2015	1	0.00000129
7	1980	1	0.00000284
8	2016	1	0.00000137
9	1989	1	0.00000156
10	1987	1	0.00000185
			# ... with 678 more rows

Cada uno de estos valores es una proporción estimada de palabras de ese documento que se generan a partir de ese tema. Por ejemplo, el modelo estima que el 0.0000009% de las palabras de la carta de 2014 se generaron a partir del tema 1.

Nótese que en la mayoría de los casos no se supera el umbral del 50% de generación para un tema concreto, sino que más bien se trata de una mezcla homogénea.

Una vez obtenidas estas probabilidades de tema, se visualiza la clasificación por aprendizaje no supervisado para la distinción de temas (Figura 50).

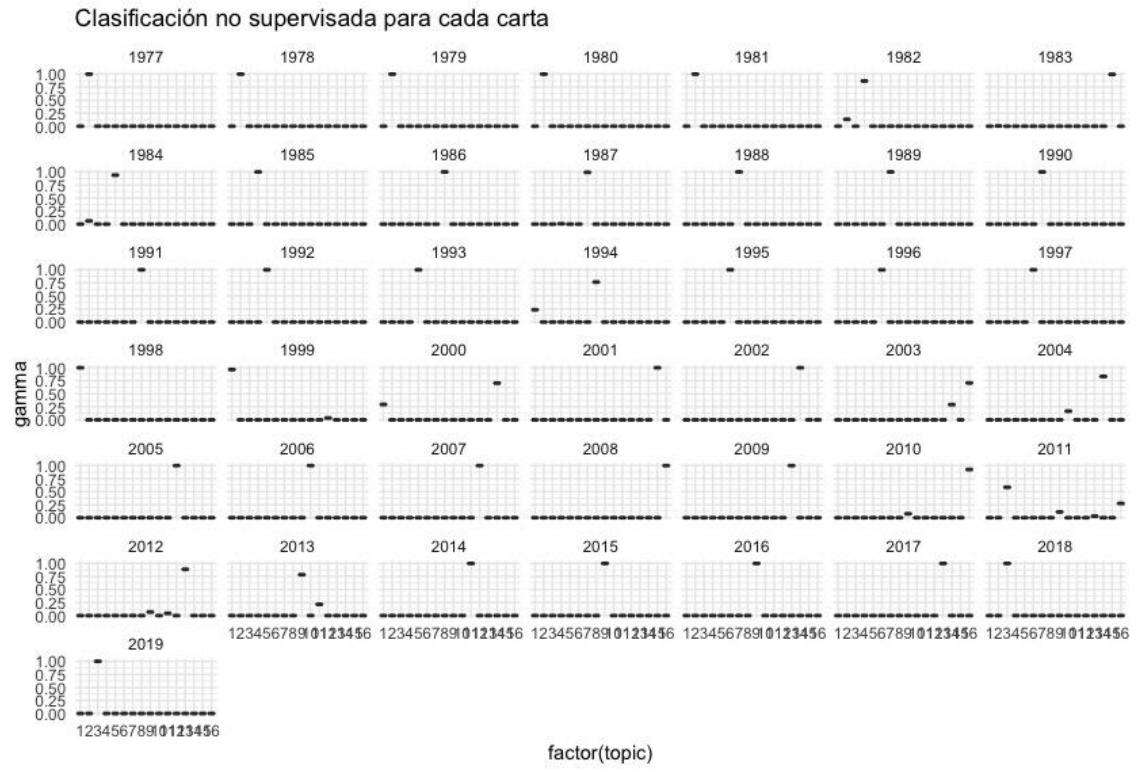


Figura 50: Las probabilidades gamma para cada tema

Se comprobó el resultado de la clasificación. Primero, se encontró el tema que estaba más asociado con cada año usando `top_n()`. Luego se comparó cada uno con el tema de "consenso" para cada año y ver cuáles se identificaron con mayor frecuencia. De entre todas las opciones probadas, 16 temas es la opción que mejor resultados permite obtener (tabla 31)

Tabla 31: Consenso

#	A tibble: 882 x 4			
	document	topic	gamma	consensus
	<chr>	<int>	<dbl>	<chr>
1	1983	1	0.507	1977
2	1983	1	0.507	1980
3	1983	1	0.507	1984
4	1983	1	0.507	1986
5	1983	1	0.507	1990
6	1983	1	0.507	1992
7	1983	1	0.507	1993
8	1983	1	0.507	1994
9	1983	1	0.507	1995
10	1983	1	0.507	2000
	# ... with 872 more rows			

7.3 Asignaciones de palabras

El siguiente paso fue asignar cada palabra de cada carta a un tema. Cuantas más palabras en un documento se asignen a ese tema, generalmente, más peso (gamma) se asignará a esa clasificación de tema del documento.

Este es el trabajo de la función `augment()`. Mientras que `tidy()` recupera los componentes estadísticos del modelo, `augment()` usa un modelo para agregar información a cada observación en los datos originales (tabla 32).

Tabla 32: Función `augment()`

	document	term	count	.topic
	<chr>	<chr>	<dbl>	<dbl>
1	2014	berkshire	203	2
2	1985	berkshire	40	2
3	1983	berkshire	39	2
4	1984	berkshire	24	2
5	1990	berkshire	28	2
6	2015	berkshire	90	2
7	1980	berkshire	33	2
8	2016	berkshire	86	2
9	1989	berkshire	24	2
10	1987	berkshire	25	2
# ... with 88,387 more rows				

Se obtuvo así un conjunto ordenados a modo de recuento agregando la columna ‘.topic’, que corresponde al tema asignado para cada término dentro de cada documento. Se destaca que la columna comienza con un punto para evitar solapar o reescribir columnas ya existentes con el mismo nombre.

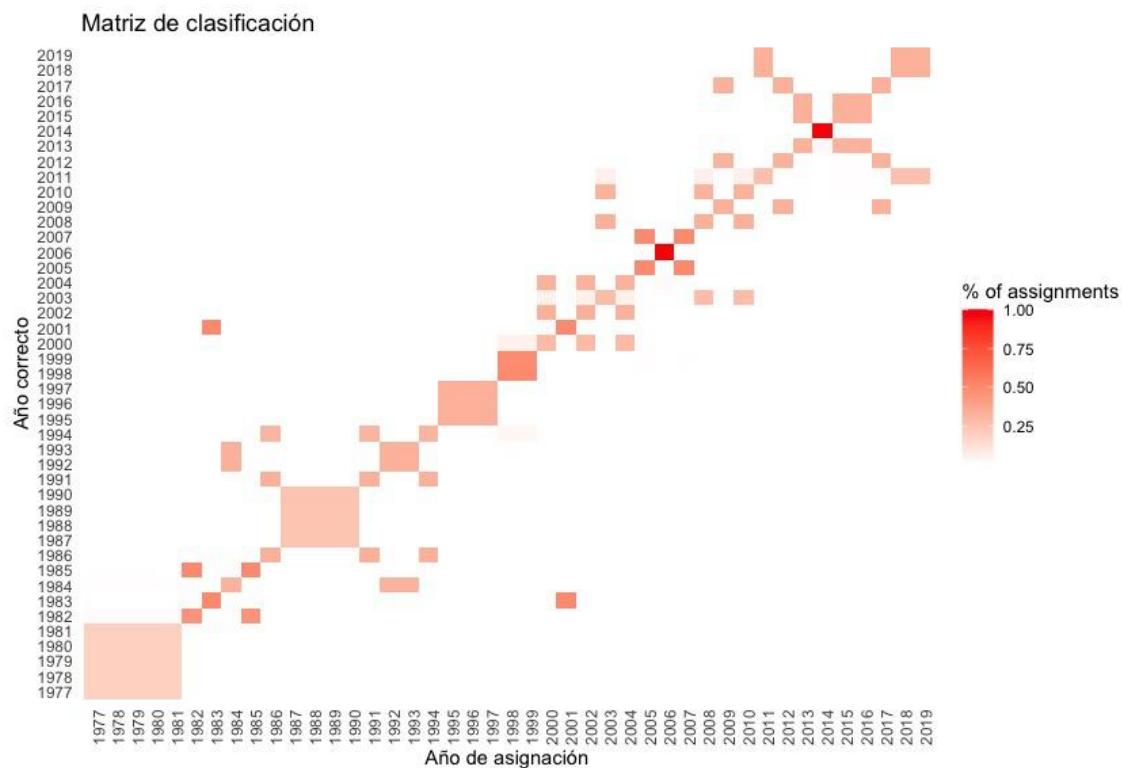


Figura 51: Matriz de clasificación

Se comprueba como muchas cartas no están correctamente clasificadas en el año correcto (nótese la escala al lado derecho: un color rojo más oscuro indicaría una mejor clasificación) (figura 51). Debemos destacar que el contenido de las mismas es muy similar y, a pesar de encontrar algunas diferencias significativas entre ellas, no son suficientes para que el algoritmo las clasifique correctamente.

8. CONCLUSIONES

El análisis de las cartas presentadas anualmente por el señor Warren Buffett durante 43 años permite entender gran parte de la intención y sentimiento que esconden sus palabras.

Se ha comprobado cómo las 20 palabras más repetidas a lo largo de los años superan la frecuencia de 600 y representan entre un 0.3% y un 1.08% del total. Estos términos económicos y financieros se distribuyen uniformemente a lo largo de los documentos sin claras excepciones. Se observó que hasta 2014 la tendencia en número de palabras empleadas en cada carta era positiva. A partir de esta fecha fue cayendo.

Gracias a la aplicación de los diferentes léxicos, se pudo comprender el sentimiento inferido detrás de las palabras. Se observó que la carta perteneciente al año 2001 arrojaba un sentimiento negativo, también la carta perteneciente al año 2008 estaba muy próxima al valor “cero”. Se pudo extraer que factores externos como las crisis financieras afectan al sentimiento general. Se pudo observar como nuevos sentimientos tenían lugar más allá del tradicional positivo o negativo.

Se realizaron análisis de 2 y 3 palabras juntas (consecutivas y no consecutivas) para interpretar el contenido de la carta e incluso el sentimiento presente. De esta forma se pudieron fijar negaciones para comprobar que conjuntos arrojaban más sentimientos positivos o negativos.

Gracias a las técnicas de frecuencia inversa se obtuvieron conclusiones acerca del contenido más relevante en cada carta. Se comprobó como de 3 a 5 palabras nos pueden revelar el tema importante que se trata en el documento.

Por último, se aplicaron técnicas de clasificación no supervisadas mediante LDA y, a pesar de que las cartas eran similares entre sí, se pudieron establecer unos parámetros de ajuste para dividir los documentos en 16 temas diferentes.

La realización de este trabajo me ha permitido entender y aprender conceptos del ámbito financiero. También me ha permitido manejar con soltura el programa R-Studio. He podido incrementar notablemente mi conocimiento en las principales librerías para la manipulación, resumen y visualizado de datos gracias a la minería de texto.

He disfrutado cada día, añadiendo líneas de código a mi trabajo e investigando nuevos algoritmos y publicaciones relacionadas. He podido adentrarme en el apasionante mundo del procesamiento del lenguaje natural.

La próxima carta del señor Warren BuffetT en este 2020 podrá confirmar que la crisis actual, no solo afecta a las acciones de Berkshire Hathaway sino al propio sentimiento detrás de las palabras de su escritor.

9. BIBLIOGRAFÍA

Loughran, T. and McDonald, B. (2011), "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance*, 66: 35-65.

Julia Silge and David Robinson (2020) "Text Mining with R - A Tidy Approach"

Michael Toth (2017) "Sentiment Analysis of Warren Buffett's Letters to Shareholders"

Susan Li (2017) "Text Mining 40 Years of Warren Buffett's Letters to Shareholders"

Juan Bosco (2018) "Análisis de sentimientos con R - Léxico Afinn"

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
<http://ggplot2.org>.

Wickham, Hadley. 2016. *tidyverse: Easily Tidy Data with ‘Spread()’ and ‘Gather()’ Functions*.
<https://CRAN.R-project.org/package=tidyr>.

Wickham, Hadley, and Romain Francois. 2016. *dplyr: A Grammar of Data Manipulation*.
<https://CRAN.R-project.org/package=dplyr>.

Font-Clos, F., Boleda, G. y Corral, Á.(2013) A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics*, 15. doi.org/10.1088/1367-2630/15/9/093033.

Montemurro, M. A. (2001). Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications* 300: 567 – 578.

