Análise e modelaxe de opinión sobre o sector enoturístico da provincia de Ourense

D. Frade-Amil, T. R. Cotos-Yáñez, M. A. Mosquera e A. Pérez-González
Departamento de Estatística e Investigación Operativa
(campus de Ourense, Universidade de Vigo)
diego.frade.amil@gmail.com, cotos@uvigo.gal, mamrguez@uvigo.gal, anapg@uvigo.gal

Resumo

O obxectivo principal deste traballo é analizar as opinións realizadas na plataforma Google Maps sobre un conxunto de adegas da provincia de Ourense. Para iso utilizaranse distintas ferramentas estatísticas que nos servirán para facer desde un estudo descritivo básico ata un estudo e unha análise do contidos dos comentarios publicados nesta plataforma. Un primeiro modelo formulado explica a valoración numérica da adega en función da lonxitude do comentario, da súa polaridade e da interacción entrambas. Tamén ao analizar o contido semántico dos comentarios se poden atopar catro grandes grupos de temas tratados: o viño, o trato recibido, aspectos da experiencia ou da visita e outro grupo de aspectos diferentes aos anteriores.

Palabras clave:

Modelaxe de opinións, *Cumulative link models*, análise de contido semántico, análise de clúster, minaría de datos

1. Introdución

O obxectivo deste subproxecto será analizar os datos extraídos previamente da plataforma Google Maps relacionados cun conxunto de adegas da provincia de Ourense. Esta tarefa dividirase en dúas partes: análise descritiva dos datos e análise das opinións. Esta segunda parte constará, á súa vez, de tres apartados:

• Análise de sentimentos. Converteremos os comentarios nun valor numérico.

- Análise da valoración. Construiremos modelos que traten de explicar a valoración en función de variables explicativas.
- Análise de contido semántico. Extraeremos os temas ou os aspectos xerais tratados nos comentarios.

Para obter estes obxectivos usaremos os paquetes e as funcionalidades que nos achega o soporte lóxico *(software)* estatístico R [1].

2. Análise descritiva

Os datos que analizamos neste subproxecto proveñen de 92 adegas das catro denominacións de orixe (en adiante, DO) localizadas na provincia de Ourense: Monterrei, Ribeira Sacra, Ribeiro e Valdeorras. Estas adegas sitúanse en 27 concellos diferentes:

		Núm. de adegas	%	Concellos	%
Denominación de orixe	Monterrei	16	17,39	4	4,81
	Ribeira Sacra	11	11,96	7	25,93
	Ribeiro	37	40,22	10	37,04
	Valdeorras	28	30,43	6	22,22
	Total	92		27	

Táboa 1. Adegas segundo a denominación de orixe e o concello

Se miramos a distribución das adegas por DO, observamos que algo máis do 40 % pertencen á DO do Ribeiro, mentres que apenas o 12 % están incluídas na DO da Ribeira Sacra. Do mesmo xeito, Ribeiro é a DO que agrupa a maior cantidade de concellos con adegas (37,04 %), fronte a Monterrei que alberga só o 14,81 % dos municipios con estes establecementos. Xeograficamente, a distribución das adegas segundo a súa DO móstrase na figura 1. De igual modo, podemos ver graficamente os concellos segundo o número de adegas localizadas neles:

Estas adegas contan nos seus perfís cun total de 1523 fotografías das súas instalacións, engadidas principalmente polos propietarios/as dos establecementos e polos autores/as das valoracións e dos comentarios. Na táboa 2 observamos que o 15,22 % das adegas carecen de fotos no seu perfil e que o 8,7 % supera o medio cento de imaxes:

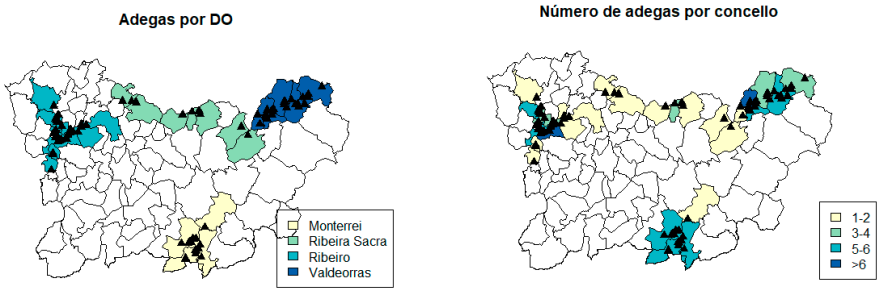

Figura 1. Adegas segundo a DO (esq.) e segundo os concellos (der.)

		Número de adegas	%
Número de fotos	0	14	15,22
	1-10	41	44,57
	11-20	12	13,04
	21-50	17	18,48
	>50	8	8,7
Total		92	

Táboa 2. Adegas segundo o número de fotos

Un total de 41 adegas (44,57 %) forman parte dalgunha ruta turística. O Ribeiro é a DO con maior número de adegas con esta condición (39,02 %) e a Ribeira Sacra, a que ten menor cantidade (14,63 %):

	Do	Monterrei	Ribeira Sacra	Ribeiro	Valdeorras	Total
Forma parte dunha ruta?	Non	9 (17,65)	5 (9,80)	21 (41,18)	16 (31,37)	51
	Si	7 (17,07)	6 (14,63)	16 (39,02)	12 (29,27)	41

Táboa 3. Adegas por DO segundo a súa pertenza a unha ruta turística

2.1. Opinións, valoracións e comentarios

Neste punto temos que facer unha pequena aclaración relativa á terminoloxía. Falaremos xenericamente de *opinións* para referirnos a todo aquilo que os usuarios e usuarias manifestaron sobre as adegas, que se compoñen de dúas partes: a *valoración* ou puntuación numérica (de 1 a 5) e o *comentario*, texto no que cada usuario ou usuaria

reflicte a súa experiencia na correspondente adega. Veremos máis adiante que hai opinións con texto e outras sen el, pero necesariamente todas as opinións contan cunha valoración. Polo tanto, estes dous termos empregaranse frecuentemente a modo de sinónimos de aquí en diante.

As opinións recompiladas sobre as adegas suman un total de 1506. Debemos ter en conta que 3 das 92 adegas mencionadas non contan con ningunha opinión. Trátase de dúas adegas da DO de Valdeorras e outra da DO da Ribeira Sacra, polo que o número de adegas en cada unha destas DO agora será de 26 e 10, respectivamente, e o total de adegas sobre o que traballaremos en diante vai ser de 89. Observamos que máis da metade das adegas (51,69 %) contan con menos de dez valoracións fronte a pouco máis do 5 % que supera o medio cento de opinións:

		Número de adegas	%
Número de opinións	0-10	46	51,69
	11-20	26	29,21
	21-50	12	13,48
	>50	5	5,62
	Total	89	

Táboa 4. Número de adegas segundo a cantidade de opinións

Se miramos o número de opinións por DO, destaca o Ribeiro con preto do 40 % das valoracións fronte a case o 11 % de opinións sobre adegas da DO de Monterrei, tal e como se recolle na táboa 5. Así mesmo, inclúese nesta táboa a densidade de frecuencia (DF), é dicir, a frecuencia relativa das opinións dividida entre o total de adegas de cada DO, e o valor máis alto corresponde á Ribeira Sacra:

		Núm. de opinións	%	Núm. de adegas	DF
DO	Monterrei	165	10,96	16	0,68
	Ribeira Sacra	435	28,88	10	2,89
	Ribeiro	600	39,84	37	1,08
	Valdeorras	306	20,32	26	0,78
Total		1506		89	

Táboa 5. Número de opinións segundo a DO

As 1506 opinións foron dadas por un total de 1371 autores/as diferentes, dos cales 545 posúen a insignia *Local Guide* (outorgada por Google aos usuarios/as de nivel 4). Do total de autores/as, 1282 (513 *Local Guide*) fixeron unha única valoración a algunha das adegas e soamente 89 (32 *Local Guide*) valoraron dúas ou máis. A puntuación ou a valoración numérica que se fai nestas opinións vén dada por unha escala de cinco valores, onde destaca o valor máximo (5) como a nota máis escollida (77,16 %):

		Número de opinións	**%**
Valoración	1	66	4,38
	2	19	1,26
	3	75	4,98
	4	184	12,22
	5	1162	77,16
Total		1506	

Táboa 6. Número de opinións segundo a valoración numérica

Así mesmo, se desagregamos as citadas puntuacións en función de posuír ou non a insignia *Local Guide*, mantense a valoración 5 como a opción maioritariamente escollida, aínda que hai unha diferenza de 10 puntos porcentuais entre os non *Local Guide* (81,16 %) e os que posúen a mencionada insignia (70,99 %):

		Ter a insignia *Local Guide*			
		Non	%	Si	%
Valoración	1	47	5,15	19	3,20
	2	13	1,42	6	1,01
	3	30	3,29	45	7,59
	4	82	8,98	102	17,20
	5	741	81,16	421	70,99

Táboa 7. Número de opinións segundo a posesión da insignia ***Local Guide***

Se atendemos ás DO, vemos que en todas elas máis do 73 % das opinións inclúen unha valoración de 5 puntos. Na táboa 8 recóllense os resultados e indícase entre parénteses a porcentaxe de opinións respecto do total por DO. Os datos da mencionada táboa 8 tamén os podemos ver de xeito gráfico (figura 2 esquerda):

		Valoracións					
		1	2	3	4	5	Total
DO	Monterrei	6 (3,64)	1 (0,61)	10 (6,06)	18 (10,91)	130 (78,79)	165
	Ribeira Sacra	21 (4,83)	6 (1,38)	17 (3,91)	71 (16,32)	320 (73,56)	435
	Ribeiro	27 (4,5)	9 (1,5)	31 (5,17)	60 (10)	473 (78,83)	600
	Valdeorras	12 (3,92)	3 (0,98)	17 (5,56)	35 (11,44)	239 (78,1)	306

Táboa 8. Número de opinións segundo a valoración e a DO

Na táboa 9 podemos ver as puntuacións segundo a adega se atope nunha ruta ou non, onde tamén observamos que a maior porcentaxe de valoración se atopa en ambos os casos para a opción 5:

		Valoracións					
		1	2	3	4	5	Total
Forma parte dunha ruta	Non	41 (5,2 %)	10 (1,2 %)	40 (5,0 %)	83 (10,5 %)	613 (77,9 %)	787
	Si	25 (3,4 %)	9 (1,2 %)	35 (4,8 %)	101 (14,0 %)	549 (76,3 %)	719

Táboa 9. Número de opinións por valoración en función da pertenza da adega a unha ruta

Con respecto ao contido da opinión, temos que facer a seguinte distinción: por unha banda están as opinións nas que só se valorou a adega, é dicir, só se indicou un valor na escala de 1 a 5 estrelas e que podemos denominar OSC (opinións sen comentario); e por outra banda, as opinións nas que a valoración numérica vai acompañada dun comentario ou dun texto no que o autor/a verte o seu parecer respecto da adega, que podemos denotar como OCC (opinións con comentario). Do total de opinións, 799 (53,05 %) carecen de texto, mentres que en 707 (46,95 %) se incluíu algún tipo de comentario, ben unha palabra, ben varias ou incluso *emojis*. Dentro deste segundo grupo, as OCC, temos as frecuencias segundo o número de palabras do texto (ou lonxitude do comentario), que se recolle na seguinte táboa:

		Número de OCC	%
Número de palabras	1-5	159	22,49
	6-10	131	18,53
	11-20	139	19,66
	21-50	185	26,17
	51-100	79	11,17
	>100	14	1,98
Total		707	

Táboa 10. Número de opinións segundo as palabras do comentario

Se disgregamos os datos da táboa anterior en función da valoración outorgada, vemos que independentemente do número de palabras que teña o comentario a valoración maioritaria é de novo 5 (na táboa 11 indícase entre parénteses a porcentaxe de opinións respecto do total por número de palabras). Pero esta situación non se dá do mesmo xeito en cada lonxitude de comentario, pois para textos con cinco ou menos palabras ou con máis de 100, a puntuación 5 foi outorgada nun 75,47 e 57,14 % das ocasións, fronte a máis do 80 % no caso de comentarios con lonxitudes intermedias (entre 6 e 100 palabras), así como no caso das opinións sen texto, nas que se outorgou a máxima puntuación no 73,59 % das veces. Tamén de xeito gráfico podemos ver o que se recolle na seguinte táboa (figura 2 dereita):

		Valoracións					
		1	2	3	4	5	Total
	0	40 (5,01)	13 (1,63)	54 (6,76)	104 (13,02)	588 (73,59)	799
	1-5	4 (2,52)	1 (0,63)	7 (4,40)	27 (16,75)	120 (75,47)	159
Número	6-10	1 (0,76)	2 (1,53)	3 (2,29)	18 (13,74)	107 (81,86)	131
de	11-20	1 (0,72)	1 (0,72)	5 (3,6)	11 (7,91)	121 (87,05)	139
palabras	21-50	9 (4,86)	2 (1,08)	4 (2,16)	20 (10,81)	150 (81,08)	185
	51-100	7 (8,86)	0 (0)	1 (1,27)	3 (3,80)	68 (86,08)	79
	>100	4 (28,57)	0 (0)	1 (7,14)	1 (7,14)	8 (57,14)	14

Táboa 11. Número de opinións segundo a valoración e o número de palabras do comentario

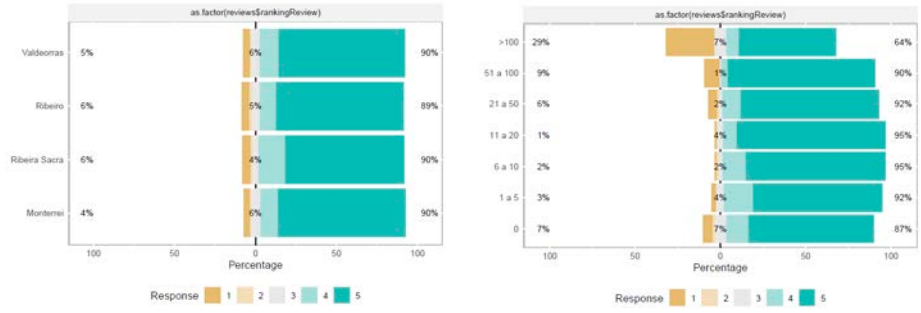

Figura 2. Valoracións por DO (esq.) e valoracións segundo a lonxitude do comentario (der.)

No que respecta ao contido dos comentarios, a análise das frecuencias de aparición de termos devólvenos que as palabras que máis veces se repiten son 'vino' e 'bodega', seguidas en terceira posición por 'visita' (táboa 12). Debemos aclarar que os termos

aparecen lixeiramente modificados da súa forma orixinal debido a un proceso de transformación previo como, por exemplo, conversión de maiúsculas a minúsculas, lematización... (estas transformacións explícanse en [2]):

Palabra	Frecuencia	Palabra	Frecuencia
vino	542	cata	102
bodega	318	buen	95
visita	206	trato	77
excelent	124	mejor	71
gracia	104	recomend	71

Táboa 12. Os dez termos máis frecuentes nos comentarios

De xeito gráfico, podemos ver mediante unha nube de palabras os 25 termos máis frecuentes, cuxos tamaño e cor dependen do seu número de aparicións (figura 3), e se deixamos a un lado os tres termos de maior frecuencia (máis de 200 aparicións), a nube das 25 palabras que máis se repiten é a que se mostra na dereita da figura 3:

Figura 3. As 25 palabras máis frecuentes (esq.) e as palabras máis frecuentes dos postos 4 ao 28 (der.)

Estas opinións extraéronse o día 5 de setembro de 2020 e constitúen unha serie temporal nun rango de nove anos previos á data de extracción. Podemos observar de xeito gráfico a evolución do número de opinións ao longo do tempo, onde o valor 0 indica que aínda non transcorreu un ano entre a data de publicación e a de extracción, e vemos que conforme nos achegamos á actualidade, o número de valoracións é cada vez maior (figura 4 esquerda):

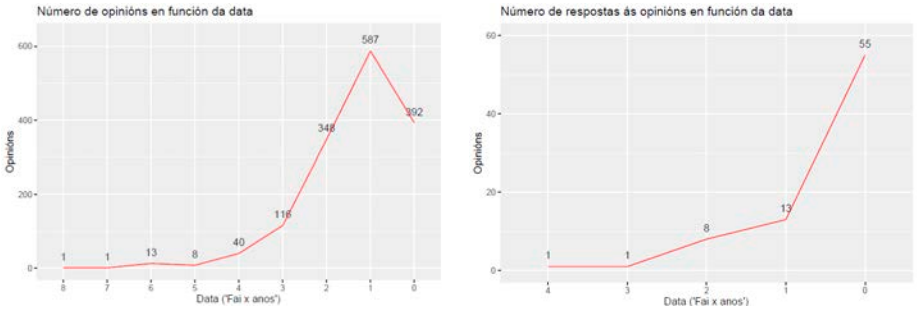

Figura 4. Data de publicación das opinións (esq.) e data da resposta ás opinións (der.)

En canto ao idioma das opinións, hai unha lixeira discrepancia entre o número de OSC (799) fronte ao número de opinións nas que non foi posible detectar o idioma (804). Tendo en conta isto, as 702 opinións nas que si se detectou o idioma (un total de 23 linguas diferentes) distribúense da maneira recollida na táboa 13 e destaca o español como o idioma preferido para escribir o comentario (74,93 %):

Idioma	Comentarios	%	Idioma	Comentarios	%
Español	526	74,93	Asturiano	6	0,85
Galego	62	8,83	Portugués	6	0,85
Aragonés	34	4,84	Catalán	5	0,71
Inglés	22	3,13	Francés	5	0,71
Italiano	8	1,14	Outros 14 idiomas	28	3,99
Total de comentarios con idioma detectado				702	

Táboa 13. Número de comentarios segundo o idioma

Un total de 583 fotografías acompañan as opinións das adegas, o que representa o 38,28 % das imaxes (1523) que conteñen os perfís destes establecementos. Pero só o 11,02 % das opinións (166) conteñen polo menos unha imaxe:

Número de fotos	0	1	2	3	4	5 a 10	>10
Número de opinións	1340	56	37	24	19	25	5

Táboa 14. Número de opinións segundo a cantidade de fotos adxuntas

Dos 707 comentarios feitos, soamente 78 recibiron algún tipo de resposta, o que supón soamente un 11 % de opinións contestadas. Graficamente, podemos ver tamén un incremento nas respostas ao longo dun período de cinco anos (figura 4 dereita). Estas respostas déronse en sete linguas diferentes e destaca o español como o idioma máis frecuente nas respostas (80,77 %):

Idioma	Número de respostas	%
Español	63	80,77
Galego	7	8,97
Neerlandés	3	3,85
Inglés	2	2,56
Aragonés	1	1,28
Alemán	1	1,28
Sueco	1	1,28
Total	78	

Táboa 15. Número de respostas por idioma

3. Análise das opinións

As opinións recollidas para este proxecto están compostas por dúas partes ben diferenciadas. Por unha banda, contan cunha parte cuantitativa, correspondente coa valoración numérica que a persoa usuaria realiza sobre o establecemento; e por outra, ten unha parte cualitativa, referida á experiencia da súa visita ás instalacións da adega. Esta experiencia vese reflectida en forma de comentario. Convén sinalar que, aínda que todas as opinións conteñen unha valoración, non necesariamente van acompañadas por un texto, feito que hai que ter en conta nas tarefas que realizaremos para a súa análise. Tales tarefas serán as seguintes:

- Análise de sentimentos. Converter o texto do comentario nunha variable cuantitativa.
- Análise da valoración. Aplicar modelos de regresión para ver o efecto de certas variables na valoración.
- Análise de contido semántico. Extraer palabras clave que nos permitan identificar a temática dos comentarios.

Como xa se comentou anteriormente, faciamos a seguinte distinción dos tipos de opinións: OSC (opinións sen comentario) e OCC (opinións con comentario). Co primeiro tipo non podemos facer ningún tipo de análise entre valoración e experiencia do usuario/a, pois carecemos desta última, polo que nas dúas tarefas arriba mencionadas nos centraremos exclusivamente nas OCC.

3.1. Análise de sentimentos

Trátase dunha técnica de minaría de datos que consiste en clasificar textos (frases, parágrafos ou documentos enteiros) de xeito automático en función da información que conteñen. Ás veces tamén se denomina *análise de contido semántico* ou *análise semántica*, pero existe unha lixeira diferenza entre ambos os conceptos, xa que se adoita usar *contido semántico* para falar de obxectos ou entidades do mundo real (obxectividade), fronte a *sentimento* para referirse ás apreciacións dos seres humanos (subxectividade) [3]. Dado que as opinións poden ter un carácter positivo ou negativo, a idea xeral da análise de sentimentos é converter o texto nun valor numérico, que denominaremos *polaridade* ou *sentimento*, termos que en diante empregaremos indistintamente. Para calcular a polaridade, bótase man do dicionario ou de listaxes de palabras positivas ou negativas, coas cales se comparan todos os termos do comentario. Estes dicionarios están formados polas palabras e por un valor numérico asociado (+1 ou -1) en función de se se considera a palabra como positiva ou como negativa, respectivamente. O procedemento consiste en atopar, para cada comentario, todos os termos que aparecen no dicionario, que se transforman no seu valor numérico asociado, cos que finalmente se calcula a polaridade ou o sentimento, valor comprendido no intervalo [-1,1]:

$$comentario \Longrightarrow \begin{pmatrix} palabra_1 \\ \dots \\ palabra_k \end{pmatrix} \Longrightarrow \begin{pmatrix} p_1 \\ \cdots \\ p_k \end{pmatrix} \Longrightarrow polaridade/sentimento$$

O soporte lóxico R dispón de varios paquetes que permiten realizar esta análise como *syuzhet*, *Rsentiment*, *SentimentAnalysis* ou *sentimentr* [4]. De entre todos, empregaremos o último por dous motivos:

- Está baseado nos monogramas, é dicir, toma cada termo do comentario e compárao co dicionario, a diferenza doutros paquetes que comparan combinacións de dous ou máis termos *(n-gramas)*.

- Ten en conta os potenciadores ou os modificadores, palabras que actúan sobre o monograma, alterando dalgún xeito o seu valor: cambio de signo (negadores), aumento ou diminución do impacto do monograma sobre o total do comentario...

Vimos na análise descritiva dos datos que o idioma predominante nas opinións era o español (táboa 13), pero tamén se recolleron comentarios noutras 22 linguas, cuxa tradución ao español empregaremos para a análise de sentimentos. Como dicionario, botamos man do léxico de opinión de Hu e Liu [5], constituído por dúas coleccións de 2005 palabras positivas e 4780 palabras negativas, e dunha listaxe de modificadores necesarios para a rutina *sentimentr* do mencionado paquete *sentimentr*. En ambos os casos, as listaxes atópanse en lingua inglesa, polo que as traducimos ao español para utilizalas, tentando arranxar na medida do posible as deficiencias da tradución automática. Por exemplo, a tradución automática do adxectivo inglés *bad* devolve a forma masculina e singular *malo* en español, obviando as demais *(mala, malos, malas)*. A mencionada rutina *sentiment()* calcula o valor numérico do comentario ou polaridade como a diferenza entre termos positivos *(#positivos)* e negativos *(#negativos)* dividida entre o total de termos «significativos» (substantivos, adxectivos, verbos, adverbios) do texto *(#all)*:

$$\frac{\#positivos - \#negativos}{\#all}$$

Como xa indicamos, este valor sitúase no intervalo [-1, 1], sendo 0 cando o número de termos positivos é igual ao de termos negativos ou cando o dicionario non detectou ningún. Ese número de termos no denominador provoca que as polaridades estean moi próximas a 0 se a cantidade de termos do comentario que aparece no dicionario é moi baixa con respecto ao número de palabras significativas do texto. Para paliar este efecto, modificamos lixeiramente a citada rutina cambiando no denominador o total de palabras pola suma de termos positivos e negativos:

$$\#positivos + \#negativos$$

Deste xeito, só se teñen en conta as palabras detectadas polo dicionario, un número inferior ao número total do comentario.

Se estimamos a densidade do sentimento en función dos distintos valores da valoración, observamos que as curvas están moi solapadas, polo que non podemos distin-

guir a que valor da puntuación se corresponde un valor calquera do sentimento (figura 5). Polo tanto, o carácter preditivo da polaridade é escaso:

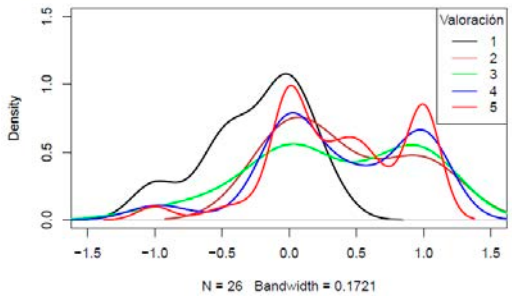

Figura 5. Densidades estimadas do sentimento en función da valoración

Os resultados obtidos na análise de sentimentos parecen verse influídos por algúns factores como poden ser a falta dun dicionario de termos e dunha listaxe de potenciadores específicos para o español, a forma de cuantificar a polaridade, determinadas características asociadas á fala (dobres negacións, ironía, sarcasmo...) ou unha posible correspondencia (directa ou inversamente proporcional) entre a lonxitude do comentario e a puntuación outorgada. A lonxitude do comentario (ou número de palabras) e a polaridade serán dúas das variables cuxo efecto sobre a valoración analizaremos no seguinte apartado.

3.2. Cumulative link models para a regresión ordinal

Sexan Y unha variable resposta ordinal, c o número de categorías de , $Y, x_1, x_2 \ldots x_p$ p-variables explicativas, o número n de elementos da mostra e h unha función *link*. Un modelo da forma:

$$h(P(Y_i \leq j|\mathbf{x_i})) = \alpha_j - \beta^{'}\mathbf{x_i} = \alpha_j - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}$$

relaciona as probabilidades acumuladas condicionadas $p(Y_i \leq j \mid \mathrm{x}_i)$ co preditor linear $\alpha_j - \beta' x_i$, para $j = 1, \ldots, c-1$ e $\mathrm{i} = 1, \ldots, n$. Os modelos desta clase denomínanse *cumulative link models* (CLM). Os coeficientes están asociados ás categorías da variable resposta e verifican $\alpha_1 < \alpha_2 < \ldots < \alpha_{(c-1)}$. O vector x_i está formado polos valores das variables explicativas e β denota o vector de coeficientes que describen os efectos destas variables. Estes últimos coeficientes son os mesmos para todos os modelos que

se constrúen variando o α_j [6]. A función *link* usada é a *logit* ou *loxística*, que toma a seguinte forma:

$$logit(P(Y_i \leq j|\mathbf{x}_i)) = \log\left(\frac{P(Y_i \leq j|\mathbf{x}_i)}{1 - P(Y_i \leq j|\mathbf{x}_i)}\right), \quad j = 1, \ldots, c-1, \ i = 1, \ldots, n.$$

Polo tanto, o noso modelo será un *cumulative logit model*, que tomará a seguinte forma xeral:

$$logit(P(Y_i \leq j|\mathbf{x}_i)) = \alpha_j - \beta'\mathbf{x}_i$$

Aínda que tamén o podemos expresar de xeito equivalente como:

$$P(Y_i \leq j|\mathbf{x}_i) = \frac{\exp(\alpha_j - \beta'\mathbf{x}_i)}{1 + \exp(\alpha_j - \beta'\mathbf{x}_i)} = \frac{1}{1 + \exp(-\alpha_j + \beta'\mathbf{x}_i)}.$$

Algunhas propiedades que verifica esta expresión son:

$$\begin{aligned} P(Y_i \leq 1|\mathbf{x}_i) &= P(Y_i = 1|\mathbf{x}_i) \\ P(Y_i = j|\mathbf{x}_i) &= P(Y_i \leq j|\mathbf{x}_i) - P(Y_i \leq j-1|\mathbf{x}_i) \\ P(Y_i = c|\mathbf{x}_i) &= 1 - P(Y_i \leq c-1|\mathbf{x}_i) \end{aligned} \tag{1}$$

A nosa variable resposta Y vai ser a valoración das adegas e contaremos con tres variables explicativas: o número de palabras do comentario que acompaña a valoración (x_1), a polaridade (ou sentimento) deste comentario (x_2) e a DO da adega valorada (x_3). Neste punto, convén facer dúas observacións:

- Tanto o número de palabras coma a polaridade ou o sentimento son variables numéricas, mentres que a DO é unha variable categórica composta de catro clases. As valoracións conforman unha escala Likert de cinco valores.
- Dado que o sentimento só se calculou para as opinións que contaban con comentario asociado, denominadas OCC (pois parece lóxico non facelo cando non hai texto para analizar), o tamaño da nosa mostra de datos redúcese de 1506 a 707. Esta variable toma valores no intervalo $[-1,1]$.
- Polo dito no punto antecedente, o número de palabras será maior ca 0. Así mesmo, dada a distribución dos datos (máis concentrados arredor de valores baixos e moi dispersos os valores altos), dános pé a realizar unha transformación logarítmica neles, contrarrestando os efectos desas respectivas concentración e dispersión. Ademais, tamén reescalamos a variable ao intervalo $[0,1]$. A pesar destas transfor-

macións, por comodidade seguiremos referíndonos a esta variable polo seu nome orixinal.

- Tras reducir o número de datos, a distribución destes en función das valoracións e das DO queda reflectida nas táboas 16 e 17, respectivamente:

		Número de OCC	%
Valoración	1	26	3,68
	2	6	0,85
	3	21	2,97
	4	80	11,32
	5	574	81,19
	Total	707	

Táboa 16. Número de OCC segundo a valoración numérica

		Número de OCC	%
DO	Monterrei	58	8,20
	Ribeira Sacra	257	36,35
	Ribeiro	282	39,89
	Valdeorras	110	15,56
	Total	707	

Táboa 17. Número de OCC segundo a DO

Ao empregar varias variables explicativas, debemos ter en conta tamén no modelo as posibles interaccións entre cada par delas. Considerando isto, a formulación do noso modelo resulta ser:

$$logit(P(Y_i \leq j|\mathrm{x}_i)) = \alpha_j - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_{12} x_{i1} x_{i2} - \beta_{13} x_{i1} x_{i3} - \beta_{23} x_{i2} x_{i3}, \qquad (2)$$

onde, $j = 1,...,4$, $i = 1,...,707$, β_1, β_2 e β_3 son os respectivos coeficientes das tres variables explicativas (tamén chamados *efectos principais*) e β_{12}, β_{13} e β_{23}, os coeficientes das súas respectivas iteracións *(efectos das iteracións)*.

O paquete *ordinal* de R permítenos aplicar regresión aos nosos datos ordinais mediante o uso de modelos CLM. *clm()* é a rutina que axusta estes modelos, que por defecto xa toma como enlace a función *logit* [7]. Tras axustar o modelo, comprobamos se

os coeficientes asociados ás variables explicativas e ás súas interaccións son significativamente distintos de 0, é dicir, contrastamos as seguintes hipóteses:

$$H_o : \beta_k = 0,\ k = 1,2,3$$

$$H_o : \beta_{kl} = 0,\ k < l,\ \ k = 1,2,\ \ l = 2,3$$

Isto verificámolo mediante a función *anova()*, que nos devolve a seguinte saída:

```
## Type II Analysis of Deviance Table with Wald chi-square tests  ##
##                          Df    Chisq    Pr(>Chisq)
##  num_palab               1   0.0787      0.77905
##  sentimento              1   1.8341      0.17564
##  DO                      3   1.4608      0.69134
##  num_palab:sentimento    1   7.3329      0.00677      **
##  num_palab:DO            3   0.8641      0.83408
##  sentimento:DO           3   0.9504      0.81325
##  ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que varios coeficientes non son significativos e o de *p-valor* máis alto é β_{13}, da interación entre o número de palabras e a DO (fila sentimento:DO). Neste punto iniciaremos un proceso de selección de variables explicativas co fin de identificar o mellor conxunto destas para axustar o modelo. A partir do modelo saturado (2), iremos extraendo en cada paso a variable ou a interacción con coeficiente menos significativo (é dicir, con *p-valor* máis alto), construíndo sucesivamente os correspondentes modelos sen as variables previamente retiradas. Na literatura especializada, este procedemento coñécese co nome de *backward stepwise selection* (selección paso a paso cara atrás) [8]. No noso caso, retiramos nun primeiro paso a mencionada interacción número de palabras-DO do modelo saturado e resulta o novo axuste na seguinte forma:

$$logit(P(Y_i \leq j|\mathrm{x}_i)) = \alpha_j - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_{12} x_{i1} x_{i2} - \beta_{23} x_{i2} x_{i3}$$

Neste novo modelo, obtemos como coeficiente menos significativo β_{23}, polo que retiramos tamén a interacción entre sentimento e DO do modelo. No seguinte axuste, que toma a forma:

$$logit(P(Y_i \leq j|\mathbf{x}_i)) = \alpha_j - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_{12} x_{i1} x_{i2},$$

volvemos comprobar se os coeficientes son significativamente distintos de , para o cal a función *anova()* nos devolve a seguinte saída:

```
## Type II Analysis of Deviance Table with Wald chi-square tests  ##
 ##                          Df    Chisq      Pr(>Chisq)
 ##   num_palab              1    0.1511      0.697483
 ##   sentimento             1    1.8920      0.168976
 ##   DO                     3    6.7515      0.080256     .
 ##   num_palab:sentimento   1    8.0710      0.004498     **
 ##   ---
## Signif. codes: 0 '***' 0.001 '**'0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vemos que o coeficiente menos significativo é β_1, seguido por β_2, correspondentes respectivamente ao número de palabras e ao sentimento, pero debemos manter estas variables no modelo atendendo ao establecido no chamado *principio xerárquico*: «Axustado un modelo con interacción, débense incluír os efectos principais aínda que o *p-valor* asociado non sexa significativo» [8]. Polo tanto, pasamos ao seguinte coeficiente menos significativo, β_3, correspondente á DO, variable que retiramos do modelo. Polo tanto, finalmente o que tomamos é o modelo formado polas variables explicativas número de palabras e sentimento, e pola interacción entre estas. Ten a seguinte forma:

$$logit(P(Y_i \leq j)|\mathbf{x_i}) = \alpha_j - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_{12} x_{i1} x_{i2}, \quad (3)$$

A función *summary()* devólvenos os datos máis relevantes dos modelos, como as estimacións dos coeficientes ou o *p-valor* asociado, entre outros:

```
## formula: valoracion ~ num_palab + sentimento + num_palab:sentimento
## data:    data
##
## link     threshold   nobs   logLik    AIC      niter    max.grad    cond.H
## logit    flexible    707    -472.94   959.87   7(2)     1.12e-13    5.8e+02
##
## Coefficients:
##                            Estimate   Std. Error   z value   Pr(>|z|)
## num_palab                  -0.6958    0.6681       -1.041    0.29766
## sentimento                 -0.6929    0.4526       -1.531    0.12584
## num_palab:sentimento        3.5749    1.1744        3.044    0.00233 **
## ---
## Signif. codes: 0 '***' 0.001 '**'0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Threshold coefficients:
##      Estimate   Std. Error   z value
## 1|2  -3.4168    0.3496       -9.773
## 2|3  -3.1951    0.3379       -9.454
## 3|4  -2.6489    0.3171       -8.354
## 4|5  -1.5823    0.2979       -5.312
```

onde observamos de novo que os coeficientes das explicativas non resultan significativos, pero si o da interacción. As estimacións de α_j, β_1, β_2 e β_{12} son:

$$\hat{\alpha}_1 = -3{,}4168,\ \hat{\alpha}_2 = -3{,}1951,\ \hat{\alpha}_3 = -2{,}6489,\ \hat{\alpha}_4 = -1{,}5823$$
$$\hat{\beta}_1 = -0{,}6958,\ \hat{\beta}_2 = -0{,}6929,\ \hat{\beta}_{12} = 3{,}5749$$

Así, substituíndo en (3) os coeficientes polas súas estimacións e incluíndo as variables correspondentes, podemos reescribir o modelo da seguinte forma:

$$logit(P(\hat{Y}_i \leq j|\mathbf{x_i})) = \hat{\alpha}_j + 0{,}6958x_{i1} + 0{,}6929x_{i2} - 3{,}5749x_{i1}x_{i2}, \quad j \in \{1,2,3,4\}, \tag{4}$$

A súa expresión equivalente como probabilidade acumulada é:

$$P(\hat{Y}_i \leq j|\mathbf{x_i}) = \frac{\exp(\hat{\alpha}_j + 0{,}6958x_{i1} + 0{,}6929x_{i2} - 3{,}5749x_{i1}x_{i2})}{1 + \exp(\hat{\alpha}_j + 0{,}6958x_{i1} + 0{,}6929x_{i2} - 3{,}5749x_{i1}x_{i2})},$$

ou simplificando:

$$P(\hat{Y}_i \leq j|\mathbf{x_i}) = \frac{1}{1 + \exp(-\hat{\alpha}_j - 0{,}6958x_{i1} - 0{,}6929x_{i2} + 3{,}5749x_{i1}x_{i2})}. \tag{5}$$

A partir de calquera das dúas fórmulas precedentes podemos calcular a probabilidade de que se asigne unha determinada valoración (de 1 a 5) en función da lonxitude do comentario e da súa polaridade. Tamén podemos escribir o modelo con interacción da seguinte forma:

$$logit(P(Y_i \leq j|\mathbf{x_i})) = \alpha_j - \beta_1 x_{i1} - (\beta_2 + \beta_{12}x_{i1})x_{i2} = \alpha_j - \beta_1 x_{i1} - \tilde{\beta}_2 x_{i2}.$$

Vemos inmediatamente que $\tilde{\beta}_2$ varía en función de , polo que axustando o valor desta variable se modificará o efecto de x_2 sobre $logit(P(Y_i \leq j|\mathbf{x_i}))$. Esta sería unha forma máis axeitada de parametrizar o modelo dado ca a variable x_2 (polaridade) se vise determinada en parte por x_1 (número de palabras). No noso caso, o modelo sería:

$$logit(P(\hat{Y}_i \leq j|\mathbf{x_i})) = \hat{\alpha}_j + 0{,}6958x_{i1} + (0{,}6929 - 3{,}5749x_{i1})x_{i2}, \quad j \in \{1,2,3,4\},$$

Outra alternativa sería escribir o modelo facendo depender x_1 de x_2, que tería a seguinte forma xeral:

$$logit(P(Y_i \leq j|\mathbf{x_i})) = \alpha_j - (\beta_1 + \beta_{12}x_{i2})x_{i1} - \beta_2 x_{i2} = \alpha_j - \tilde{\beta}_1 x_{i1} - \beta_2 x_{i2}.$$

Ao substituír os coeficientes estimados, o modelo é:

$$logit(P(\hat{Y}_i \leq j|\mathbf{x_i})) = \hat{\alpha}_j + (0{,}6958 - 3{,}5749x_{i2})x_{i1} + 0{,}6929x_{i2}$$

Na figura 6 podemos ver de xeito gráfico o efecto da interacción entre as variables explicativas. En cada subfigura represéntase a probabilidade de obter a correspondente valoración en función das mencionadas variables. O uso das cores do arco da vella que proporciona a rutina *image()* permite mostrar as diferentes probabilidades sen necesidade de realizar unha representación en tres dimensións, que sería o habitual pero cuxo gráfico implica unha interpretación laboriosa. Así, vemos franxas de cores correspondentes ás áreas cunha determinada probabilidade, delimitadas polas *curvas de nivel*, liñas que unen os puntos para os que o modelo predí a mesma probabilidade de asignarlle a correspondente valoración. Inclúense tamén os datos da nosa mostra, sinalados como puntos negros:

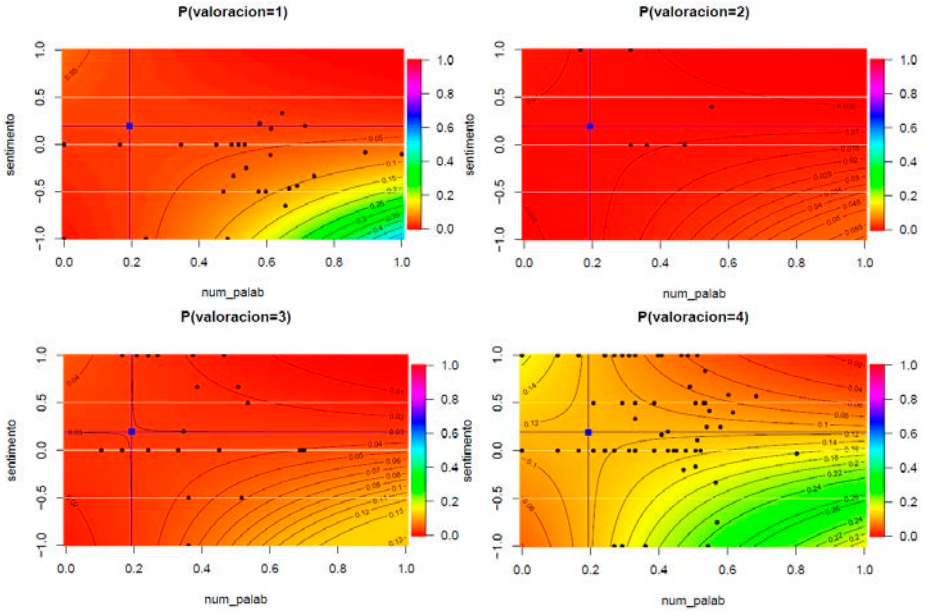

Figura 6. Probabilidade de ter valores 1, 2, 3 ou 4

Ademais, pódense distinguir catro rexións determinadas e delimitadas entre si por dúas rectas perpendiculares (representadas na figura en cor azul). Estas rectas calcúlanse como os puntos nos cales se anulan os coeficientes $\tilde{\beta}_1$ e $\tilde{\beta}_2$:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_{12} x_{i2} = 0 \Leftrightarrow r: x_{i2} = -\frac{\hat{\beta}_1}{\hat{\beta}_{12}}$$

$$\tilde{\beta}_2 = \hat{\beta}_2 + \hat{\beta}_{12} x_{i1} = 0 \Leftrightarrow s: x_{i1} = -\frac{\hat{\beta}_2}{\hat{\beta}_{12}}$$

O punto no que se intersecan estas rectas r e s resulta ser un *punto de sela* ou *punto minimax*. Este é un tipo de punto sobre unha superficie caracterizado por ter pendente cero e tomar un valor máximo ou un valor mínimo segundo se considere unha dirección ou a súa ortogonal. Denotarémolo por (sinalado por un cadrado azul na figura 6) e, substituíndo os coeficientes do modelo, toma o seguinte valor:

$$\text{ps} = (ps_1, ps_2) = \left(-\frac{\hat{\beta}_2}{\hat{\beta}_{12}}, -\frac{\hat{\beta}_1}{\hat{\beta}_{12}}\right) = (0{,}1938; 0{,}1946)$$

Dado que a variable «número de palabras» transformouse da forma que xa indicamos, se desfacemos tales transformacións é sinxelo comprobar que o valor $ps_1 = 0{,}1938$ se correspondería cun comentario de 3,6249 palabras, marcando fronteira entre comentarios con tres ou menos palabras e comentarios con catro ou máis palabras. Unha vez identificadas estas rectas e a súa intersección, podemos ver cal é o efecto das variables e da interacción nelas. Tomando, por exemplo, un punto (ps_1, x_{i2}) na recta s, temos:

$$\begin{aligned} -\hat{\beta}_1 ps_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_{12} ps_1 x_{i2} &= \hat{\beta}_1 \frac{\hat{\beta}_2}{\hat{\beta}_{12}} - \hat{\beta}_2 x_{i2} + \hat{\beta}_{12} \frac{\hat{\beta}_2}{\hat{\beta}_{12}} x_{i2} = \frac{\hat{\beta}_1 \hat{\beta}_2}{\hat{\beta}_{12}} - \hat{\beta}_2 x_{i2} + \hat{\beta}_2 x_{i2} \\ &= \frac{\hat{\beta}_1 \beta_2}{\hat{\beta}_{12}}. \end{aligned} \tag{6}$$

De igual modo sucede para un punto (x_{i1}, ps_2) da recta r e, como é lóxico, para ps. Polo tanto, o impacto das variables na resposta para puntos destas rectas vén dado polo produto dos efectos principais dividido polo efecto da interacción. A partir de ps, definimos as catro rexións disxuntas seguintes:

$$R_1 = \{\mathbf{x_i} = \frac{x_{i1}, x_{i2}}{x_{i1}} < 0{,}1938, x_{i2} < 0{,}1946\}$$

$$R_2 = \{\mathbf{x_i} = \frac{x_{i1}, x_{i2}}{x_{i1}} > 0{,}1938, x_{i2} < 0{,}1946\}$$

$$R_3 = \{\mathbf{x_i} = \frac{x_{i1}, x_{i2}}{x_{i1}} < 0{,}1938, x_{i2} > 0{,}1946\}$$

$$R_4 = \{\mathbf{x_i} = \frac{x_{i1}, x_{i2}}{x_{i1}} > 0{,}1938, x_{i2} > 0{,}1946\}$$

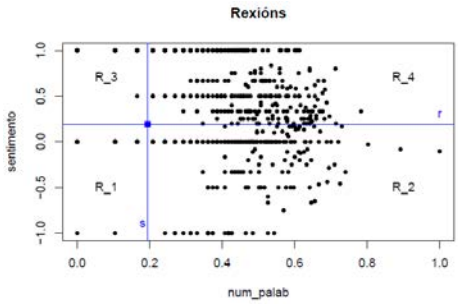

Figura 7. Gráfico das rexións, incluíndo as rectas que as determinan (liñas azuis), o punto de sela (cadrado azul) e os puntos da mostra (puntos negros)

É dicir, as rexións R_1 e R_4 están constituídas polos puntos cuxas dúas coordenadas son simultaneamente menores e maiores ca as de ps, respectivamente. En R_2, a primeira coordenada é maior ca ps_1 e a segunda, menor ca ps_2, mentres que en R_3 sucede o contrario. Noutras palabras e dito a grandes trazos, a rexión R_1 representa os comentarios con poucas palabras e a polaridade neutra ou negativa, mentres que esta mesma polaridade e un maior número de palabras caracterizan os comentarios que caen en R_2; de xeito análogo, R_3 recolle os comentarios con poucas palabras e polaridade positiva, e en R_4 recaen os comentarios con maior número de palabras e polaridade tamén positiva. Convén indicar que os datos da nosa mostra se distribúen de xeito moi diferente entre as parellas de rexións $R_2 - R_4$ (que aglutinan sobre o 87%) e $R_1 - R_3$ (onde se concentran en moi poucos valores distintos). Dada a baixa densidade de datos nestas últimas dúas rexións, posiblemente o modelo devolva resultados pouco fiables ou incoherentes, motivo polo cal debemos ser cautelosos á hora de interpretalos. Así, se lle botamos unha ollada á figura 6, vemos (grosso modo) que as rexións R_2 e R_3 (especialmente a primeira) acollen as probabilidades máis altas de asignar unha valoración entre 1 e 4; mentres que no caso de ter unha valoración 5, as maiores probabilidades danse para os puntos das rexións R_1 e R_4. Polo dito anteriormente, cabe pensar que os resultados para R_1 e R_3 resulten enganosos, algo que discutiremos polo miúdo máis adiante.

Para ilustrar o comportamento do modelo, tomaremos un punto de cada unha das catro rexións xunto co punto , que empregaremos para comparar, e calcularemos para cada un deles a probabilidade de obter como valoración as distintas posibilidades (1, 2, 3, 4 ou 5). Os puntos elixidos son:

$$\mathbf{p_1} = (0{,}1;\ -0{,}5) \in R_1,\ \mathbf{p_2} = (0{,}8;\ -0{,}5) \in R_2,$$
$$\mathbf{p_3} = (0{,}1;\ 0{,}7) \in R_3,\ \mathbf{p_4} = (0{,}8;\ 0{,}7) \in R_4$$

Os valores da primeira coordenada destes puntos $(0{,}1$ e $0{,}8)$ corresponderíanse, desfacendo as transformacións, con comentarios de 1,9435 e 203,5828palabras respectivamente, que podemos interpretar que en realidade se tratan de senllas opinións de 2 e 204 palabras. Botando man de (4) e (6), para o punto de sela *ps* temos:

$$logit(P(\hat{Y} \leq j|\mathbf{ps})) = \hat{\alpha}_j + \frac{\hat{\beta}_1\hat{\beta}_2}{\hat{\beta}_{12}} = \hat{\alpha}_j + \frac{0{,}6958 \cdot 0{,}6929}{3{,}5749} = \hat{\alpha}_j + 0{,}1349,$$

de onde sacamos que (por [5]):

$$P(\hat{Y} \leq j|\mathbf{ps}) = \frac{1}{1 + \exp(-\hat{\alpha}_j - 0{,}1349)}$$

Usando as propiedades da probabilidade acumulada (véxase [1]), vemos por exemplo que a probabilidade de que a *ps* se lle outorgue como valoración un 1 é:

$$P(\hat{Y} = 1|\mathbf{ps}) = P(\hat{Y} \leq 1|\mathbf{ps}) = \frac{1}{1 + \exp(-\hat{\alpha}_1 - 0{,}1349)} =$$
$$= \frac{1}{1 + \exp(3{,}4168 - 0{,}1349)} = 0{,}0362,$$

mentres que a probabilidade de que obteña un 5 como valoración resulta ser:

$$P(\hat{Y} = 5|\mathbf{ps}) = 1 - P(\hat{Y} \leq 4|\mathbf{ps}) = 1 - \frac{1}{1 + \exp(-\hat{\alpha}_4 - 0{,}1349)} == 1 - \frac{1}{1 + \exp(1{,}5823 - 0{,}1349)}$$
$$= 0{,}8097$$

Daquela, é máis probable que a se lle asigne unha valoración de 5 puntos, como tamén podemos comprobar graficamente na figura 6. Agora repetimos o proceso cos puntos de cada rexión. Para facilitar a lectura, soamente realizaremos os cálculos para , $\mathbf{p_1} = (0{,}1;\ -0{,}5)$, xa que para os demais as operacións son similares, e recolleremos todos os resultados na táboa 18. Para o citado punto, substituíndoo en (4), temos:

$$logit(P(\hat{Y} \leq j|\mathbf{p_1})) = \hat{\alpha}_j + 0{,}6958 \cdot 0{,}1 - 0{,}6929 \cdot 0{,}5 + 3{,}5749 \cdot 0{,}1 \cdot 0{,}5 = \hat{\alpha}_j - 0{,}0981$$

A probabilidade de que este punto reciba unha valoración de 1 punto é:

$$P(\hat{Y} = 1|\mathbf{p_1}) = P(\hat{Y} \leq 1|\mathbf{p_1}) = \frac{1}{1 + \exp(3{,}4168 + 0{,}0981)} = 0{,}0289$$

e de que teña unha valoración de 5 puntos:

$$P(\hat{Y} = 5|\mathbf{p_1}) = 1 - P(\hat{Y} \leq 4|\mathbf{p_1}) = 1 - \frac{1}{1 + \exp(1{,}5823 + 0{,}0981)} = 0{,}843$$

Á vista da táboa 18, observamos que se os valores das variables son simultaneamente máis altos ou máis baixos ca os das coordenadas de ps (rexións R_4 e R_1, respectivamente), a probabilidade de obter unha valoración 1 (2, 3 ou 4) redúcese, mentres que a probabilidade de obter un 5 como valoración vai en aumento.

	Puntos x				
	ps (0,1938; 0,1946)	**p1** (0,1; –0,5)	**p2** (0,8; –0,5)	**p3** (0,1; 0,7)	**p4** (0,8; 0,7)
$P(\hat{Y} = 1\|\mathbf{x})$	0,0362	0,0289	0,1447	0,0426	0,0124
$P(\hat{Y} = 2\|\mathbf{x})$	0,0086	0,0069	0,0297	0,0100	0,0030
$P(\hat{Y} = 3\|\mathbf{x})$	0,0301	0,0244	0,0928	0,0349	0,0109
$P(\hat{Y} = 4\|\mathbf{x})$	0,1156	0,0968	0,2473	0,1304	0,0466
$P(\hat{Y} = 5\|\mathbf{x})$	0,8097	0,8443	0,4855	0,7821	0,9271

Táboa 18. Probabilidade de obter valoracións de 1 a 5

Por outra parte, cando o valor dunha variable aumenta con respecto da correspondente coordenada de ps ao tempo que na outra variable sucede o contrario (rexións R_2 e R_3), a probabilidade de obter como valoración un 1 (2, 3 ou 4) é cada vez maior, mentres que a de obter un 5 é pouco a pouco máis baixa. Nesta ollada xeral que nos presentan as probabilidades destes puntos, podemos intuír algúns resultados que non parecen fiables, como xa adiantamos anteriormente. Para facilitar a análise, podemos considerar dúas situacións xerais, atendendo ás parellas de rexións que mencionamos no seu momento:

- Comentarios con poucas palabras (R_1 – R_3). Estariamos ante usuarios/as que poderiamos cualificar de breves e de concisos. A súa visión da adega pode ser positiva, que dará lugar a unha polaridade positiva (R_3), ou negativa, que resultará nunha polaridade máis negativa (R_1). Agora ben, no primeiro caso cabería esperar que unha alta polaridade se traducise con gran probabilidade en valoracións altas, en contraposición con valoracións máis baixas. O modelo así o reflicte por exemplo no caso da valoración 5, pero as probabilidades non son tan altas como se podería esperar e van a menos conforme o número de palabras redúcese e a polaridade alcanza o seu valor máximo. No segundo caso (en R_1), deberiamos esperar con

gran probabilidade valoracións baixas, que serían acordes co carácter negativo do comentario. Pero o modelo devólvenos todo o contrario, é dicir, con maior probabilidade un comentario neste suposto recibiría un 5 como valoración. Neste caso vemos acentuado o efecto da baixa densidade de datos nesta rexión que «adultera» os resultados do modelo.

- Comentarios con maior número de palabras ($R_2 - R_4$). Os usuarios/as reflicten a súa experiencia nun texto máis ou menos longo, onde inclúen os aspectos positivos ou negativos que consideren oportunos. Unha experiencia negativa pode dar lugar a un comentario longo no cal o usuario/a trate de xustificarse explicando o seu parecer, que tenderá a conter termos que xeren unha polaridade negativa (R_3), e ir acompañado por unha valoración baixa. O modelo recolle acertadamente este razoamento e o 1 resulta a valoración máis probable canto maior é o número de palabras e máis negativa a polaridade. E, aínda que habitualmente poida parecer que unha boa experiencia se resume de forma breve e concisa (véxase o caso de R_2), tamén pode darse o caso de que os usuarios/as que a percibiron deste xeito o queiran mostrar nun texto máis extenso, que tenderá a ter termos que dean pé a unha polaridade positiva (R_4), e ir acompañado dunha valoración alta. O modelo tamén o mostra así, asignando maiores probabilidades á valoración 5 para comentarios neste suposto.

Ante o visto, os resultados que devolve o modelo para a rexión R_1carecen de toda lóxica, polo que poderemos desbotalos. Así, como conclusión, os resultados extraídos deste modelo (4) permítennos afirmar que para comentarios cuxa polaridade sexa positiva, a probabilidade de obter unha valoración alta será cada vez maior conforme aumente o número de palabras do texto, mentres que se o comentario sobe de catro palabras e a polaridade é negativa, as valoracións baixas serán máis probables segundo aumente o número de palabras.

3.3. Análise de contido semántico

Baixo esta epígrafe trataremos de ver que se esconde nos textos das opinións. Por exemplo, extraeremos as palabras máis importantes (ou *palabras clave*) e, a partir delas, trataremos de distinguir que temas ou asuntos se recollen nos comentarios. Para esta análise, botaremos man principalmente dos seguintes tres paquetes de R:

- *Udpipe*. Permite a *tokenización* dun texto, isto é, a división deste nas súas partes máis elementais (palabras, símbolos e signos de puntuación); a etiquetaxe das partes da fala (POS, *parts of speech*) (substantivos, verbos, adxectivos...), a lematización das palabras e a análise de dependencia entre elas [9].
- *Textrank*. Permite a identificación das frases máis relevantes dun texto e das palabras relevantes ou palabras clave [10].
- *Wordnet*. Proporciona unha interface á base de datos léxica do idioma inglés *WordNet* da Universidade de Princeton (EUA), que permite agrupar palabras en conxuntos baseados en relacións léxicas (sinónimos, antónimos, hiperónimos, hipónimos...) [11][12].

Dado que este último paquete só está dispoñible para a lingua inglesa, vímonos na obriga de traducir os comentarios das opinións a esta lingua e realizar a análise con este idioma como base, manténdonos igualmente na mesma cantidade de comentarios: 707.

3.3.1. Tokenización, lematización e etiquetaxe POS

As funcións do paquete *udpipe* precisan dun modelo para poder executarse. Estes modelos están baseados en bancos de árbores de dependencias universais, que están dispoñibles para máis de 65 idiomas [9]. No caso da lingua inglesa, hai nove posibles modelos para elixir, dos cales tomamos o *UD_English-EWT*, que resulta ser un dos máis completos nas distintas utilidades que os conforman (*tokenización* e segmentación de palabras, morfoloxía e sintaxe). A rutina *udpipe_annotate()* toma como atributos de entrada este modelo e os comentarios, devolvendo como saída un *data.frame*. De todas as columnas deste conxunto de datos, resaltaremos as seguintes catro: *doc_id*, refírese ao comentario; *sentence_id*, identifica cada frase de cada comentario (os puntos «.» separan as frases); *lemma*, o lema de cada palabra de cada frase/comentario; e *upos*, a etiqueta POS correspondente a ese lema (substantivo, adxectivo, verbo, signo de puntuación...). Na táboa 19 resúmense os datos correspondentes ás distintas etiquetas POS, identificadas coas respectivas categorías gramaticais:

Etiqueta POS	Categoría		Etiqueta POS	Categoría	
ADJ	Adxectivo	2034	PART	Partícula	355
ADP	Aposición	1790	PRON	Pronome	1897
ADV	Adverbio	1087	PROPN	Nome propio	761

Etiqueta POS	Categoría		Etiqueta POS	Categoría	
AUX	Verbo auxiliar	852	PUNCT	Signos de puntuación	2428
CCONJ	Conxunción coordinada	935	SCONJ	Conxunción subordinada	244
DET	Determinante	2230	SYM	Símbolo	46
INTJ	Interxección	46	VERB	Verbo	1903
NOUN	Nome/Substantivo	4042	X	Outros	7
NUM	Numeral	247			
		Total	**20904**		

Táboa 19. Elementos dos comentarios segundo a categoría gramatical

3.3.2. Palabras clave

As rutinas do paquete *textrank* permiten identificar as frases máis relevantes e as palabras clave dos comentarios [10]. Debido ao número de comentarios (que dá lugar a un maior número de frases), non buscaremos cales son as frases máis relevantes, e centraremos a nosa análise nas palabras clave. Para atopar estes termos máis salientables, botamos man da función *textrank_keywords()* do mencionado paquete. Esta rutina crea unha rede de palabras, construída tendo en conta que palabras se acompañan unhas a outras. Créase un enlace entre dúas palabras se unha segue a outra, cuxo peso se incrementará canto maior sexa o número de veces que aparecen estas palabras seguidas. Dadas estas conexións entre palabras, aplícase o algoritmo *PageRank* para obter a importancia de cada termo. *Textrank_keywords()* toma como atributo de entrada os *lemma* obtidos con anterioridade e dános a opción de escoller como relevantes unha ou varias categorías gramaticais. Por defecto, toma como palabras clave un terzo das que forman a rede. Tras realizar unha serie de tarefas de depuración (principalmente corrección de erros ortográficos e de tradución), calculamos as palabras clave indicando nun primeiro momento como categorías relevantes substantivos («NOUN») e adxectivos («ADJ»). Así, pódense ver que combinacións de palabras destas categorías son máis frecuentes, por exemplo:

```
##      keyword          ngram freq
##  9   good-wine        2     60
## 16   excellent-wine   2     39
## 31   family-winery    2     26
## 36   wine-taste       2     22
## 38   quality-wine     2     20
```

```
## 40  great-wine          2   20
## 49  white-wine          2   16
## 53  excellent-treatment 2   15
## 66  different-wine      2   12
## 68  beautiful-winery    2   11
## 69  good-treatment      2   11
```

Observamos que o termo *wine* (viño) vai acompañado con frecuencia de adxectivos de cualificación positiva *(good*, *excellent*, *great...)*. Pero para tratar de ver a temática dos comentarios, interésannos unicamente os substantivos como palabras clave. Entón, na rutina *textrank_keywords* indicamos só esta categoría como relevante e que nos devolva a totalidade das palabras clave. Tras unha depuración previa de termos non desexados (en xeral, malas traducións), obtemos unha listaxe final de 153 palabras clave. As palabras clave aparecen ordenadas segundo o índice de importacia que calcula internamente a rutina *textrank_keywords()*, como mencionamos anteriormente. Neste punto, debemos aclarar que, debido aos procesos de tradución e de lematización, a frecuencia de aparicións dalgúns dos termos aquí recollidos pode non coincidir co seu respectivo valor da táboa 12 da análise descritiva.

3.3.3. Temática dos comentarios

As palabras claves extraídas poden indicarnos que temas ou aspectos se están tratando nos comentarios. Para identificalos, usaremos o paquete *wordnet*, tratando de agrupar as palabras claves en función da súa temática. Como xa indicamos, este paquete proporciona unha interface á base de datos léxica do inglés *WordNet* da *Princeton University* [11]. Nela, recóllense catro categorías gramaticais (substantivos, verbos, adxectivos e adverbios) agrupadas en conxuntos de *sinónimos cognitivos* ou *synsets*, que expresan un determinado concepto, os cales están conectados mediante relacións semánticas e léxicas. A relación máis frecuente entre *synsets* é a relación *super-subordinada*, consistente en agrupar conceptos máis xerais (por exemplo, automóbil) con outros máis específicos (ambulancia, autobús, caravana...). Debemos establecer nesta relación a seguinte nomenclatura:

- *Hiperónimo*. Clase á que pertence un determinado conxunto de termos. Por exemplo, «automóbil» é un hiperónimo de «ambulancia».
- *Hipónimo*. Palabra que pertence a unha clase máis xeral. Por exemplo, «autobús» é un hipónimo de «automóbil».

Así, esta relación tamén se coñece como *hiperonimia* ou *hiponimia*. Esta relación é transitiva, isto é, se unha ambulancia é un tipo de automóbil e un automóbil é un tipo de vehículo a motor, daquela unha ambulancia é un tipo de vehículo de motor. Esta relación establece unha estrutura xerárquica para as palabras, que ten como categoría inicial ou nodo raíz a clase «entidade» *(entity)*.

O paquete *wordnet* proporciona unha serie de filtros para realizar as buscas na base de datos –*getFilterTypes()*–, de entre os cales eliximos o de atopar o termo exacto. Unha vez tomado o filtro, seleccionamos os termos que queremos con base nunha ou en varias das categorías gramaticais –*getIndexTerms()*– e, a continuación, escollemos os *synsets* nos que se atopan os termos –*getSynsets()*– [12]. Como no noso caso queremos establecer os temas ou os aspectos que tratan os comentarios, quedarémonos cos *synsets* que nos devolvan hiperónimos das palabras clave, coa función *getRelatedSynsets()*. Dado que cun só hiperónimo non obtemos conexión entre as palabras clave, repetimos varias veces este procedemento, empregando como termos para buscar os sucesivos hiperónimos atopados. Este procedemento empregámolo xa no seu momento para depurar as palabras clave, descartando os termos para os cales non se detectou hiperónimo ningún. Tras atopar algunha conexión entre palabras clave, representamos as relacións de hiperonimia mediante un diagrama de árbore, para o que botamos man dos paquetes *data.tree*, *treemap* e *DiagrammeR* para configurar a estrutura xerárquica da árbore, visualizala e representala nun gráfico, respectivamente (na figura 8 só representamos as relacións das dez primeiras palabras clave). Como xa mencionamos anteriormente, a categoría raíz é *entity*, de onde parten as demais clases. Indicamos mediante puntos suspensivos («...») que existen outras categorías entre a raíz e a do hiperónimo «máis grande» identificado para cada palabra clave:

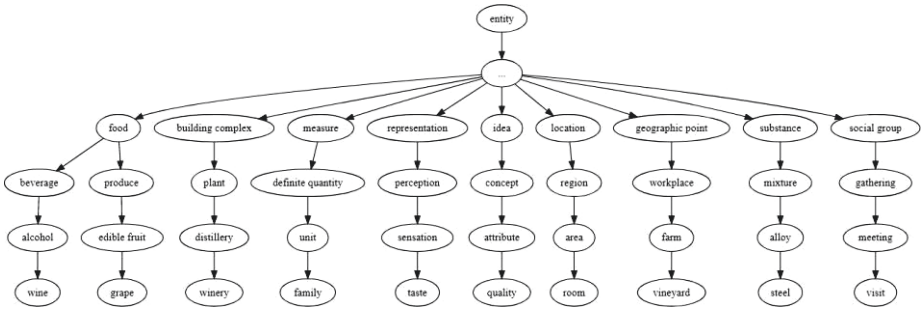

Figura 8. Árbore xerárquica das palabras clave

Pero nesta forma de agrupar as palabras clave non tivemos en conta as seguintes dúas cuestións:

- Tomamos de cada vez un único hiperónimo, descartando outros do mesmo nivel, que poderían ser máis axustados ao contexto da palabra clave. Estamos, pois, perdendo ou descartando posible información relevante.
- As palabras clave poden non estar ao mesmo nivel do nodo raíz e pode darse o caso de que unha palabra clave englobe outras varias.

Con este segundo punto en mente, tratamos de realizar a clasificación das palabras clave comezando por *entity*, de onde xorde unha árbore cos sucesivos niveis dados polos hipónimos (véxase a figura 9 para os dous niveis):

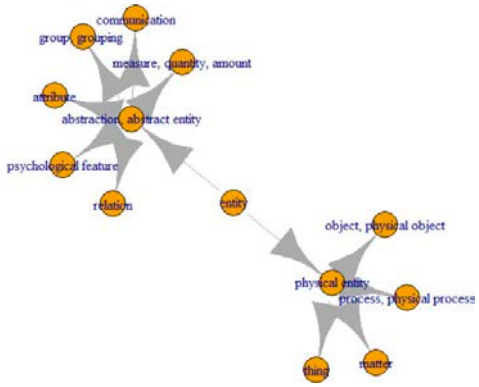

Figura 9. Árbore xerárquica do nodo raíz e dous niveis de hipónimos

Se facemos o corte no nivel dous e consideramos as categorías correspondentes aos hipónimos resultantes, cada unha destas estará composta por un número determinado de palabras clave, recollido na seguinte táboa:

	Palabras clave
Attribute	11
communication	9
group, grouping	6
matter	11
measure, quantity, amount	11
object, physical object	57
process, physical process	2
psychological feature	32

	Palabras clave
relation	4
thing	1
Total	144

Táboa 20. Número de palabras clave nas categorías xeradas polo nivel 2 de hipónimos de *entity*

Para tratar de atopar a temática dos comentarios, botamos man da *análise clúster* ou *clustering*. Esta técnica consiste en dividir o conxunto de datos en grupos (clústeres), de xeito que os membros dun mesmo grupo sexan similares entre si e entre elementos de distintos grupos haxa algunha diferenza [8]. Como a priori non sabemos o número de posibles temas tratados nos comentarios, escollemos o *clustering xerárquico*, que nos proporciona unha representación gráfica das observacións en forma de árbore (denominada *dendograma*). Este gráfico é útil á hora de decidir o número de clústeres que podemos tomar. Isto marca a diferenza entre o *clustering* xerárquico e o *clustering por k-medias*, onde se necesita establecer previamente un número de grupos.

Para realizar o *clustering*, construímos tantas variables coma categorías temos e cunha lonxitude igual ao número de comentarios (707). Serán variables indicadoras, isto é, tomarán o valor 1 cando algunha palabra clave da categoría correspondente aparece no comentario e 0 cando isto non ocorra:

$$V_{ij} = \begin{cases} 1 & \text{se } keyword.cat_j \in comentario_i \\ 0 & \text{noutro caso} \end{cases}, \quad i = \{1\ldots707\},\ j = \{1\ldots153\}$$

Dado que son variables binarias, a diferenza entre os elementos (comentarios) debémola calcular cunha métrica acorde, que indicaremos co parámetro *method* = «binary» na función *dist()* de R. Creamos así a matriz de *disimilitude* (non semellanza), que precisamos para realizar o *clustering* coa función *hclust()*. O dendograma resultante vémolo na esquerda da figura 10:

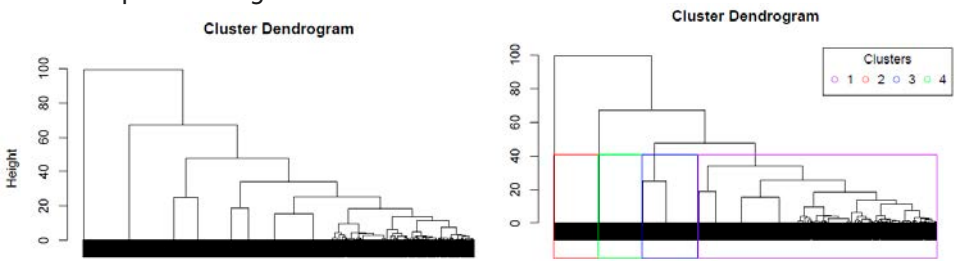

Figura 10. Dendograma (esq.) e dendograma cos clústeres (der.)

Podemos probar a tomar catro grupos ou clústeres, realizando outros tantos cortes no dendograma coa función *cutree()* (que tamén permite realizar os cortes indicando a altura á que os queremos). Tras realizar os cortes, obtemos o dendograma da dereita da figura 10. Para estes catro grupos calculamos a frecuencia de aparición das palabras clave (véxase a táboa 21). Á vista da táboa, o clúster 2 contén comentarios sen palabras clasificadas en ningunha das categorías, mentres que os demais clústeres conteñen comentarios con palabras encadradas en polo menos unha categoría. Convén destacar os clústeres 3 e 4, onde no 100 % dos comentarios temos palabras das categorías *psychological feature* (característica psicolóxica) e *matter* (materia), respectivamente, que poderían dar lugar a dous temas sobre os que se está a falar nos comentarios deses respectivos grupos. Agora ben, se observamos as palabras clave agrupadas en cada unha destas categorías, vemos que non existe unha relación temática aparente entre elas. Por exemplo, *wine* e *grape* están na categoría *matter*, mentres que *winemaking* o está en *psychological feature*, pero poderían referirse a un tema común: «viño». Así mesmo, neste último conxunto temos termos como *taste*, *wedding* ou *prevention* sen que compartan un tema común máis alá do mencionado conxunto ao que pertencen. Tras realizar un certo número de probas, podemos concluír que incrementar o número de clústeres non soluciona este inconveniente, así como tampouco o fai se consideramos niveis máis baixos de hipónimos de *entity*, onde se dá a mesma situación. Ademais, esta segunda opción leva un procedemento moi laborioso debido á gran cantidade de categorías posibles que van xurdindo e á limitación das funcionalidades do paquete *wordnet* para realizar tal tarefa:

	C1	C2	C3	C4
attribute	27,66	0,00	0,00	0,00
communication	12,93	0,00	0,00	0,00
group, grouping	14,74	0,00	0,00	0,00
matter	67,12	0,00	58,25	100,00
measure, quantity, amount	12,47	0,00	0,00	0,00
object, physical object	82,99	0,00	0,00	0,00
process, physical process	0,45	0,00	0,00	0,00
psychological feature	67,35	0,00	100,00	0,00
relation	7,48	0,00	0,00	0,00
thing	3,40	0,00	0,00	0,00

Táboa 21. Porcentaxe de aparición das categorías (> 10 %) nos grupos

Para solucionar este inconveniente, optamos por realizar o *clustering* tomando directamente como base as propias palabras clave, nun proceso análogo ao anterior. O dendograma resultante para estes termos móstrase na esquerda da figura 11. Á vista da figura, podemos probar a crear catro clústeres realizando outros tantos cortes no dendograma coa función *cutree()* (que tamén permite realizar os cortes indicando a altura á que os queremos). Logo de realizar os cortes, obtemos o dendograma da dereita da figura 11. Para estes catro grupos calculamos a frecuencia de aparición das palabras clave. Na seguinte táboa mostramos as palabras con máis dun 10 % de aparicións en polo menos un dos clústeres.

Palabra clave	Clúster			
	1	2	3	4
wine	63,23	0,00	77,27	100,00
winery	44,84	0,00	0,00	0,00
taste	27,02	0,00	0,00	0,00
quality	11,07	0,00	0,00	0,00
vineyard	11,07	0,00	0,00	0,00
visit	34,71	0,00	0,00	0,00
treatment	10,69	0,00	100,00	0,00
spectacular	10,88	0,00	0,00	0,00

Táboa 22. Porcentaxe de aparición das palabras clave (> 10 %) nos grupos

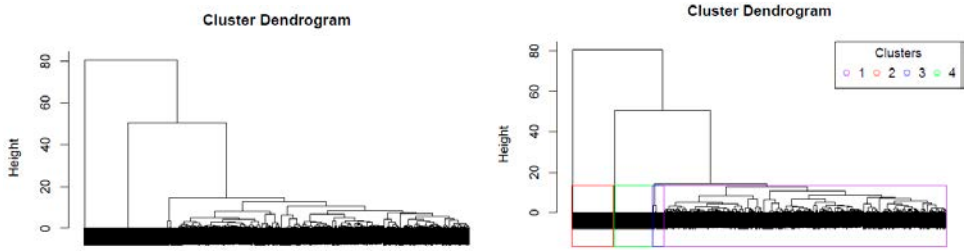

Figura 11. Dendograma (esq.) e dendograma cos catro clústeres (der.)

Ao observar os resultados, o clúster 4 está definido polos comentarios que exclusivamente se refiren ao viño *(wine)*, así como co terceiro grupo, onde o 100 % dos comentarios aluden ao trato *(treatment)* recibido na adega. Os restantes *clústeres* son casos contrapostos: mentres que o primeiro o conforman comentarios con polo menos unha das palabras clave (que podemos resumir baixo a temática «experiencia xeral»,

por exemplo), no grupo 2 recaen os comentarios que tratan aspectos máis aló dos definidos polas palabras clave ou temas residuais. En resumo, as palabras clave permítennos extraer como temas que se tratan nos comentarios: o viño, o trato recibido, aspectos varios da experiencia e outros. Convén indicar que, se elevamos o número de clústeres, obteranse outros resultados, mais a súa interpretación pode ser máis difícil.

4. Conclusións

No presente documento damos conta da análise de datos extraídos de Google relativos a 1506 opinións deixadas por persoas usuarias de 92 adegas da provincia de Ourense. Centramos a nosa atención nas opinións conformadas por valoración numérica e por comentario (un total de 707). Estes comentarios transformáronse nun valor numérico que nos deu a súa polaridade (negativos ou positivos). Comprobamos o escaso poder preditivo da polaridade, ao non poder distinguir a que valoración se corresponde un valor calquera da polaridade. Con todo, empregando esta variable e o número de palabras de cada comentario, puidemos construír un modelo de regresión ordinal que nos permite explicar a valoración en función da lonxitude do comentario, da polaridade e da interacción entrambas. Así, para comentarios con polaridade positiva, a probabilidade de obter unha valoración alta aumenta ao incrementar o número de palabras do texto, mentres que se o comentario ten máis de catro palabras e a polaridade é negativa, as valoracións baixas serán as máis probables. E finalmente, analizando a temática dos comentarios, obtivemos catro grandes grupos de temas tratados polos usuarios e usuarias: o viño, o trato recibido, aspectos da experiencia ou da visita e outro grupo de aspectos diferentes aos anteriores.

5. Referencias

[1] Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[2] Calvo Torres M (2017). Text Analytics para Procesado Semántico (Traballo de fin de máster, Universidade de Vigo). http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1 475.pdf

[3] Qi Z, Storey VC, Jabr W (2015). Sentiment Analysis Meets Semantic Analysis: Constructing Insight Knowledge Bases.

[4] Naldi M (2019). A review of sentiment computation methods with R packages. arXiv preprint arXiv:1901.08319.

[5] Hu M, Liu B (2004). Mining and summarizing customer reviews. En: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.

[6] Agresti A (2003). Categorical data analysis (Vol. 482). John Wiley & Sons.

[7] Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the R package ordinal. Submitted in J. Stat. Software.

[8] James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning with Applications in R, Springer.

[9] Wijffels J (2020). UDPipe Natural Language Processing - Text Annotation. https:// cran.r- project.org/web/packages/udpipe/vignettes/udpipe-annotation.html.

[10] Wijffels J (2020). Textrank for summarizing text. https://cran.r-project.org/web/ packages/t extrank/vignettes/textrank.html.

[11] Princeton University (2010) About WordNet. https://wordnet.princeton.edu/. Princeton University, New Jersey (EUA).

[12] Feinerer I (2017). Introduction to the wordnet Package. https://cran.r-project.org/ web/pack ages/wordnet/vignettes/wordnet.pdf

Proxectos INOU 2020

Investigación aplicada na provincia de Ourense

Coordinadora:

de Blas Varela, Esther

DEPUTACION OURENSE

Vicerreitoría do Campus de Ourense

Universidade Vigo

Proxectos INOU 2020. Investigación aplicada na provincia de Ourense

Autores/as:
Comendador Rey, Beatriz Pilar
Braña Rey, Fátima
González Rodríguez, Rubén
González Rufino, María Encarnación
Rivas Siota, Sandra
Gullón Estévez, Beatriz
Rodríguez Campo, María Lorena
Pavón Rial, María Reyez
Mosquera Rodríguez, Manuel Alfredo

Coordinadora:
de Blas Varela, Esther

Comisión de Avaliación:
García Señorán, María del Mar
Franco Matilla, María Inmaculada
Prada Rodríguez, Julio
Rodeiro Iglesias, Javier
Méndez Penín, Arturo Jose
Sampayo Fernández, Jose A.

Vicerreitoría do Campus de Ourense-Campus Auga
Universidade de Vigo
Ourense, 2021

Nº de páxinas: 238

ISBN: 978-84-8158-916-0

Edición
Vicerreitoría do Campus de Ourense - Campus Auga
www.uvigo.gal/campus/ourense-campus-auga

Maquetación
Rodi Artes Gráficas, S. L.

Índice