

Sentiment Analysis

Clasificación de Textos

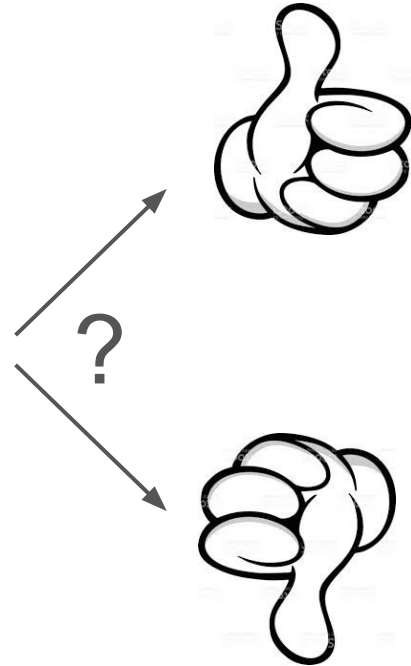
Detección de sentimiento

@juanma25



Que bien que estoy!

Empanadas + Netflix , infalible!



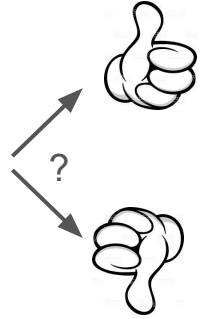
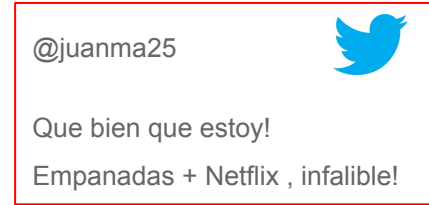
Métodos:

Opción 1

- Uso de reglas y expresiones regulares

Opción 2

- Supervised Machine Learning



Sentiment Lexicons

Bing Liu Opinion Lexicon

- 2006 positive words: accessible, easy, gifted, ingenious, ingeniously
- 4783 negative words: abnormal, darkness, friction, zombie...

Características:

- Errores ortográficos, lunfardo, variantes morfológicas
- Orientado a opiniones de productos

Sentiment Lexicons

Bing Liu Opinion Lexicon

- 2006 positive words: accessible, easy, gifted, ingenious, ingeniously
- 4783 negative words: abnormal, darkness, friction, zombie...

@juanma25



Que **bien** que estoy!

Empanadas + Netflix , **infalible**!

Sentiment Lexicons

Bing Liu Opinion Lexicon

- 2006 positive words: accessible, easy, gifted, ingenious, ingeniously
- 4783 negative words: abnormal, darkness, friction, zombie...

- Positive = 3/7

- Negative = 0/7

@juanma25



Que **bien** que estoy!

Empanadas + Netflix , **infalible**!

Sentiment Lexicon

Hedonometer (10k palabras)

word	Happiness
Laughter	8.50
Happiness	8.44
love	8.42
...	
rape	1.44
suicide	1.30
terrorist	1.30

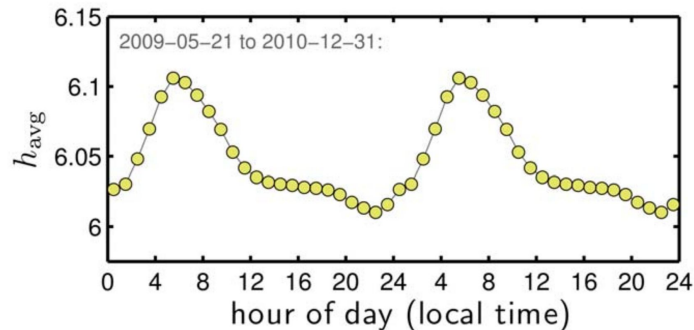
6.1

Groucho Marx:

Outside of a dog, a book is a man's best friend. Inside of a dog it's too dark to read.

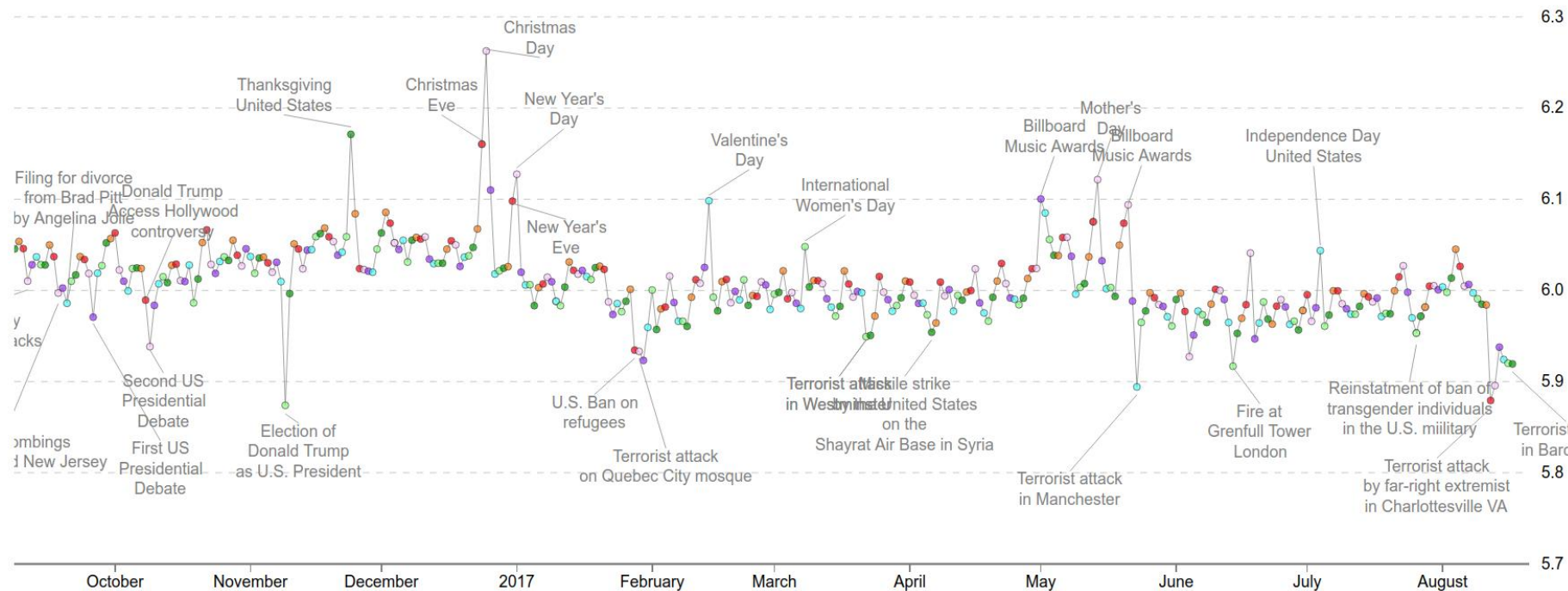
7.2 5.0 6.9 5.1 5.0 5.9

Felicidad promedio = 5.09

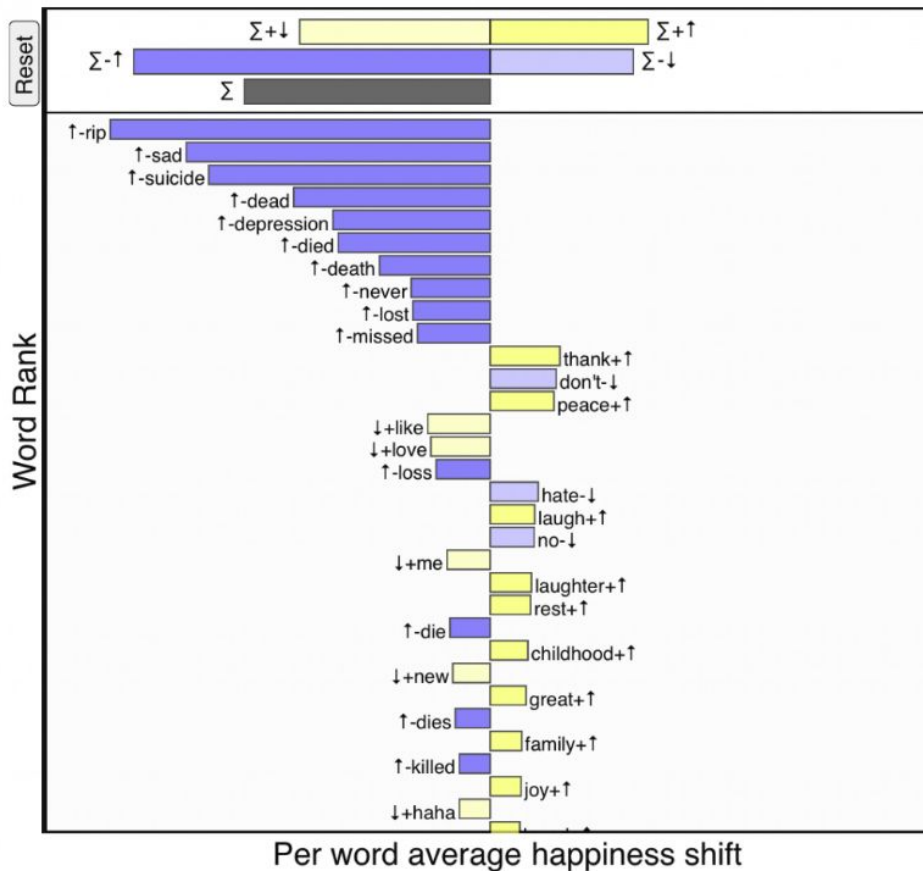


Léxico de felicidad para twitter

Hedonometer



Palabras representativas en el día de la muerte de Robin Williams



Word shift showing how happiness dropped on Twitter for the day of Robin Williams's death compared to the previous seven days. Click for interactive version.

Más allá del sentimiento

- General Inquire (182 listas, abierto para investigación)
 - Positiv, Negativ
 - Strong, Weak
 - Active, Passive
 - Pleasure, Pain
 - Arousal (excitation)
 - Content: Academ, Econ@, Relig, etc.
 - ...

Stone et al. (1966) The General Inquirer: A Computer Approach to Content Analysis.

<http://www.wjh.harvard.edu/~inquirer/homecat.htm>

Más allá del sentimiento

- General Inquire (182 listas, abierto para investigación)
- LIWC
 - Linguistic Dimensions (pronouns, Articles, Negations)
 - Psychological Processes (Affective processes, Anger, pos/neg emotions)
 - Social processes (Family, Friends)
 - Cognitive processes (Insight, Causation, Certainty)
 - Biological processes (Health, sexual, Ingestion, Body)
 - Time orientations (Past, Present, Future)
 - ...

Más allá del sentimiento

- General Inquire (182 listas, abierto para investigación)
- LIWC
- NRC EMOLEX
 - Anger - fear - anticipation - trust - surprise - sadness - joy - disgust
 - Positive - Negative

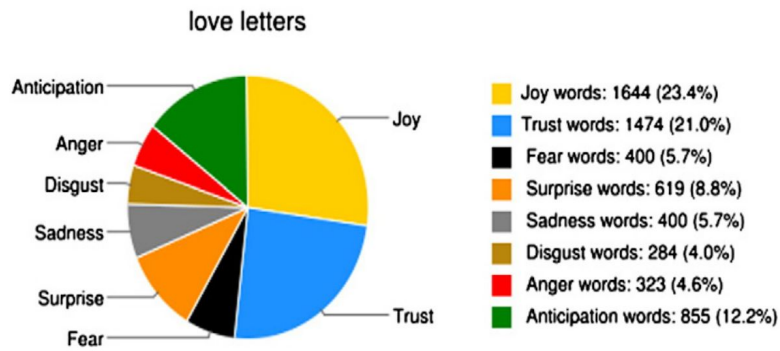


Fig. 4. Percentage of emotion words in the love letters corpus.

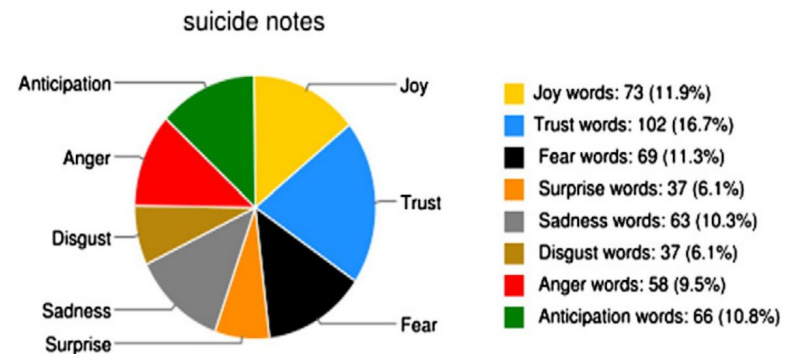


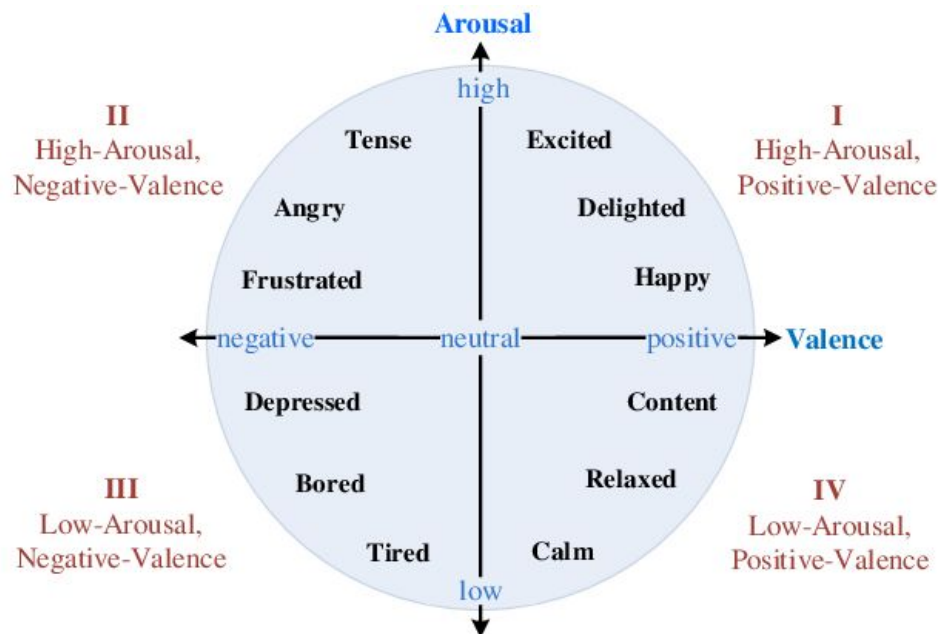
Fig. 6. Percentage of emotion words in the suicide notes corpus.

Más allá del sentimiento

- General Inquire (182 listas, abierto para investigación)
- LIWC
- NRC EMOLEX
- MPQA Subjectivity lexicon
 - positive vs negative
 - objective vs subjective

Más allá del sentimiento

- General Inquire (182 listas, abierto para investigación)
- LIWC
- NRC EMOLEX
- MPQA Subjectivity lexicon
- Warriner's Norms
 - Valence (pos-neg)
 - Arousal (Intensidad)
 - Dominance



Más allá del sentimiento

- General Inquire (182 listas, abierto para investigación)
- LIWC
- NRC EMOLEX
- MPQA Subjectivity lexicon
- Warriner's Norms
- Concreteness ratings

Más allá del sentimiento

- General Inquire (182 listas, abierto para investigación)
- LIWC
- NRC EMOLEX
- MPQA Subjectivity lexicon
- Warriner's Norms
- Concreteness ratings
- Dictionary of Affect in Language (DAL)
 - Pleasantness
 - Activation
 - Imagery (qué tan fácil es de imaginar)

En español

- Spanish Dictionary of Affect in Language (SDAL)
 - Pleasantness
 - Activation
 - Imagery (qué tan fácil es de imaginar)

En español

- Spanish Dictionary of Affect in Language (SDAL)
- Spanish Warriner's Norms
 - Valence
 - Arousal

En español

- Spanish Dictionary of Affect in Language (SDAL)
- Spanish Warriner's Norms
- Emociones (Ferré):
 - Felicidad
 - Enajo
 - Miedo
 - Disgusto
 - Tristeza

En español

- Spanish Dictionary of Affect in Language (SDAL)
- Spanish Warriner's Norms
- Emociones (Ferré):
- The Madrid Affective Database for Spanish (MADS)
 - Familiaridad (que tan cotidiana es la palabra)
 - Dominancia: (que tanto dominio uno tiene. ej: la ira es una emoción dominante, mientras que el miedo es una emoción sumisa)
 - Edad de adquisición (estimación de edad de adquisición)
 - Experiencia sensorial (que tan sensorial es)

¿Cómo se arma un Lexicon?

- Anotaciones humanas



¿Cómo se arma un Lexicon?

- Anotaciones humanas



- De manera semi-supervisada (semantic orientation, Turney)

1. Defino seeds:

neg_words = [mala, malo, horrible, fea, feo...]

pos_words = [buena, bueno, hermosa, hermoso, linda, lindo...]

¿Cómo se arma un Lexicon?

➤ Anotaciones humanas



➤ De manera semi-supervisada (semantic orientation, Turney)

1. Defino seeds:

neg_words = [mala, malo, horrible, fea, feo...]

pos_words = [buena, bueno, hermosa, hermoso, linda, lindo...]

2. Defino una métrica de cercanía semántica

¿Cómo se arma un Lexicon?

- Anotaciones humanas



- De manera semi-supervisada (semantic orientation, Turney)

1. Defino seeds:

neg_words = [mala, malo, horrible, fea, feo...]

pos_words = [buena, bueno, hermosa, hermoso, linda, lindo...]

2. Defino una métrica de cercanía semántica

3. Para cada palabra del vocabulario calculo su orientación (sentimiento)

$$polaridad(w) = \sum_{w_p \in pos_words} cercania(w, w_p) - \sum_{w_n \in neg_words} cercania(w, w_n)$$

Semantic Orientation (Turney 2001)

$$polaridad(w) = \sum_{w_p \in pos_words} PMI(w, w_p) - \sum_{w_n \in neg_words} PMI(w, w_n)$$

Turney usa bigramas

Pos tag 1	Pos tag 2
adjetivos	sustantivo
adverbio	verbo
adverbio adjetivos sustantivo	adjetivos (no a la iz de un sustantivo)

Semantic Orientation (Turney 2001)

$$\begin{aligned} \textit{polaridad}(w) &= PMI(w, w_p) - PMI(w, w_n) \\ &= \log \left(\frac{p(w, w_p)}{p(w) \cdot p(w_p)} \right) - \log \left(\frac{p(w, w_n)}{p(w) \cdot p(w_n)} \right) \end{aligned}$$

Diagram illustrating the Semantic Orientation formula. The word w_p is associated with the positive sentiment "Excellent" (green arrow), and the word w_n is associated with the negative sentiment "Poor" (red arrow).

Semantic Orientation (Turney 2001)



$$\begin{aligned} \textit{polaridad}(w) &= \textit{PMI}(w, w_p) - \textit{PMI}(w, w_n) \\ &= \log \left(\frac{p(w, w_p)}{p(w).p(w_p)} \right) - \log \left(\frac{p(w, w_n)}{p(w).p(w_n)} \right) \\ &= \log \left(\frac{p(w, w_p).p(w_n)}{p(w_p).p(w, w_n)} \right) \end{aligned}$$

Excellent

Poor

Semantic Orientation (Turney 2001)

$$\begin{aligned} polaridad(w) &= PMI(w, w_p) - PMI(w, w_n) \\ &= \log \left(\frac{p(w, w_p)}{p(w).p(w_p)} \right) - \log \left(\frac{p(w, w_n)}{p(w).p(w_n)} \right) \\ &= \log \left(\frac{p(w, w_p).p(w_n)}{p(w_p).p(w, w_n)} \right) \\ &= \log \left(\frac{p(w|w_p)}{p(w|w_n)} \right) \end{aligned}$$

Excellent  Poor 

Probabilidad condicional
 $p(A, B) = p(A|B)p(B)$

Como lo calculó? (Turney 2001)

$$\begin{aligned} polaridad(w) &= PMI(w, w_p) - PMI(w, w_n) \\ &= \log \left(\frac{p(w, w_p) \cdot p(w_n)}{p(w_p) \cdot p(w, w_n)} \right) \\ &= \log \left(\frac{\frac{f(w, w_p)}{k \cdot N} \cdot \frac{f(w_n)}{N}}{\frac{f(w_p)}{N} \cdot \frac{f(w, w_n)}{k \cdot N}} \right) \end{aligned}$$

Como lo calculó? (Turney 2001)

$$\begin{aligned} polaridad(w) &= PMI(w, w_p) - PMI(w, w_n) \\ &= \log \left(\frac{p(w, w_p) \cdot p(w_n)}{p(w_p) \cdot p(w, w_n)} \right) \\ &= \log \left(\frac{\frac{f(w, w_p)}{k \cdot N} \cdot \frac{f(w_n)}{N}}{\frac{f(w_p)}{N} \cdot \frac{f(w, w_n)}{k \cdot N}} \right) \\ &= \log \left(\frac{f(w, w_p) \cdot f(w_n)}{f(w_p) \cdot f(w, w_n)} \right) \end{aligned}$$

**Usa un buscador para
calcular las f's**

**La polaridad de un review es la
polaridad promedio de sus
bigramas**

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
little difference	JJ NN	-1.615
clever tricks	JJ NNS	-0.040
programs such	NNS JJ	0.117
possible moment	JJ NN	-0.668
unethical practices	JJ NNS	-8.484
low funds	JJ NNS	-6.843
old man	JJ NN	-2.566
other problems	JJ NNS	-2.748
probably wondering	RB VBG	-1.830
virtual monopoly	JJ NN	-2.050
other bank	JJ NN	-0.850
extra day	JJ NN	-0.286
direct deposits	JJ NNS	5.771
online web	JJ NN	1.936
cool thing	JJ NN	0.395
very handy	RB JJ	1.349
lesser evil	RBR JJ	-2.288
Average Semantic Orientation		-1.218

Pre-procesamiento

- Tokenización casual:
 - NLTK TweetTokenizer o WWBP happierfuntokenizing*
 - '@juancapo95: :-) Mejor día de mi vidaaaaaa!! #quesoydulce <3'
 - [':', ':-)', 'mejor', 'día', 'de', 'mi', 'vida', '!', '!', '#quesoydulce', '<3']

*<https://github.com/dlatk/happierfuntokenizing>

Pre-procesamiento

- Propagación de negaciones
 - ['Esto', 'no', 'está', 'bueno', 'ni', 'rico', '.', 'En', 'la', '...']
 - ['Esto', 'no', 'no_está', 'no_bueno', 'no_ni', 'no_rico', '.', 'En', 'la', '...']

Pre-procesamiento

- Propagación de negaciones
 - ['Esto', 'no', 'está', 'bueno', 'ni', 'rico', '.', 'En', 'la', '...']
 - ['Esto', 'no', 'no_está', 'no_bueno', 'no_ni', 'no_rico', '.', 'En', 'la', '...']
- Otros:
 - Postags
 - Ortografía
 - Stopwords
 - Normalización de fechas, números, lugares, productos, etc..

Métodos totalmente supervisados

Medidas de Personalidad (Big-5)

- **Openness**
Abierto a nuevas experiencias
- **Conscientiousness**
Disciplina, organización
- **Extroversion**
Grado de sociabilidad
- **Agreeableness**
Que tan amistosa, agradable, cooperativa es
- **Neuroticism**
Grado de estabilidad emocional



75.000 usuarios de Facebook
tomaron un test de personalidad

Las palabras y la personalidad

Preprocesamiento

➤ happyfuntokenizer

Las palabras y la personalidad

Preprocesamiento

- happyfuntokenizer
- bigramas y trigramas con $pmi > 2 * len$

len={2,3} según si toma
bigrama o trigrama

Phrase = n-grama

$$pmi(phrase) = \log \frac{p(phrase)}{\prod_{w \in phrase} p(w)}$$

Las palabras y la personalidad

Preprocesamiento

- happyfuntokenizer
- bigramas y trigramas con $pmi > 2 * len$
- Distribución de tokens de cada sujeto

$len=\{2,3\}$ según si toma
bigrama o trigrama

Phrase = n-grama

$$pmi(phrase) = \log \frac{p(phrase)}{\prod_{w \in phrase} p(w)}$$

$$p(phrase | subject) = \frac{freq(phrase, subject)}{\sum_{phrase' \in vocab(subject)} freq(phrase', subject)}$$

Las palabras y la personalidad

Preprocesamiento

- happyfuntokenizer
- bigramas y trigramas con $pmi > 2 \cdot len$
- Distribución de tokens de cada sujeto
- transformacion de Anscombe

$len=\{2,3\}$ según si toma
bigrama o trigrama

Phrase = n-grama

$$pmi(phrase) = \log \frac{p(phrase)}{\prod_{w \in phrase} p(w)}$$

$$p(phrase | subject) = \frac{freq(phrase, subject)}{\sum_{phrase' \in vocab(subject)} freq(phrase', subject)}$$

$$p_{ans}(phrase | subject) = 2\sqrt{p(phrase | subject) + 3/8}$$

Regresión Lineal

➤ Para cada test fiteo:

$$y = w_0 + \sum_{i \in [1, N]} w_i \cdot p(\text{phrase}_i)$$

Diagram illustrating the linear regression formula for fitting a test:

- y : Valor obtenido en el test (Value obtained in the test)
- w_0 : Bias term
- $\sum_{i \in [1, N]}$: Sum over all n-grams in the test
- w_i : Peso de cada n-grama en el test (Weight of each n-gram in the test)
- $p(\text{phrase}_i)$: Distribución de n-gramas (Distribution of n-grams)

Regresión Lineal

Extroversion

n-grama	Peso (w)
party	0.110
cant wait	0.099
...	...
internet	-0.066
anime	-0.087

Agreeableness

n-grama	Peso (w)
excited	0.059
a great	0.056
...	...
shit	-0.110
fuck	-0.12

Con los w_i calculo los Big-5 de personalidad de un nuevo documento!

$$y = w_0 + \sum_{i \in [1, N]} w_i \cdot p(\text{phrase}_i)$$

Las palabras y la personalidad

Elementos de bienestar (PERMA)

- Positive emotions
- Engagement
- Relationships
- Meaning
- Accomplish



Anotaciones manuales de perfiles de facebook de 5100 personas

A. Schwartz et al. predicting individual well-being through the language of social media

Comparación de Corpus



Will never pay to see Star wars again

This Film, was so off in so many ways, as a Star Wars movie Fan (not books) you will wonder what they were smoking. You will also feel insulted that the Characters you once knew are bent out of shape in odd ways. If you cant sit thourgh it twice you will literally start mothing "why" at loads of scenes.....

Star Wars:
The Last Jedi (2017)



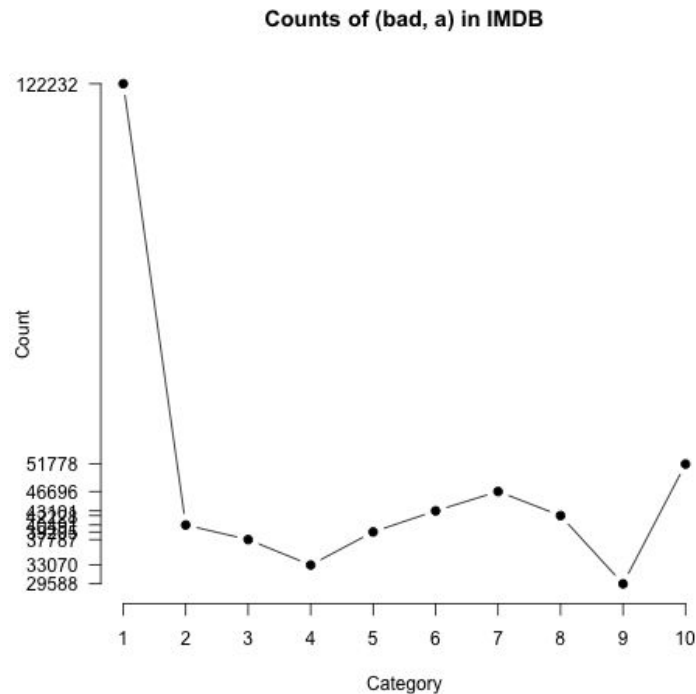
A fresh new beginning with the potential for a new direction and ideas. Loved it.

The Last Jedi instilled the same excitement, awe and wonder I first felt when seeing A New Hope back in 1977.

I agree with Luke and Kylo. It's time to forget the past and to move forward....

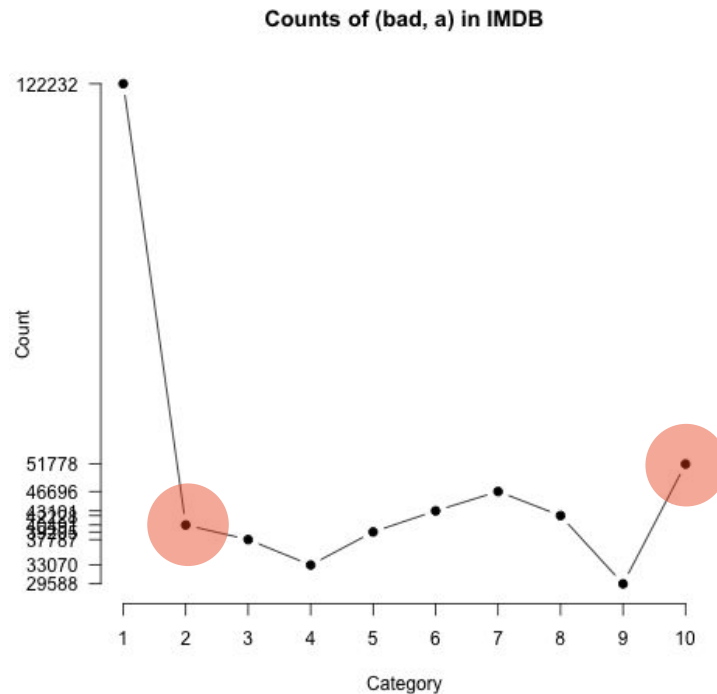
Las palabras y el sentimiento

¿Cómo se distribuye la palabra “BAD” en las críticas?

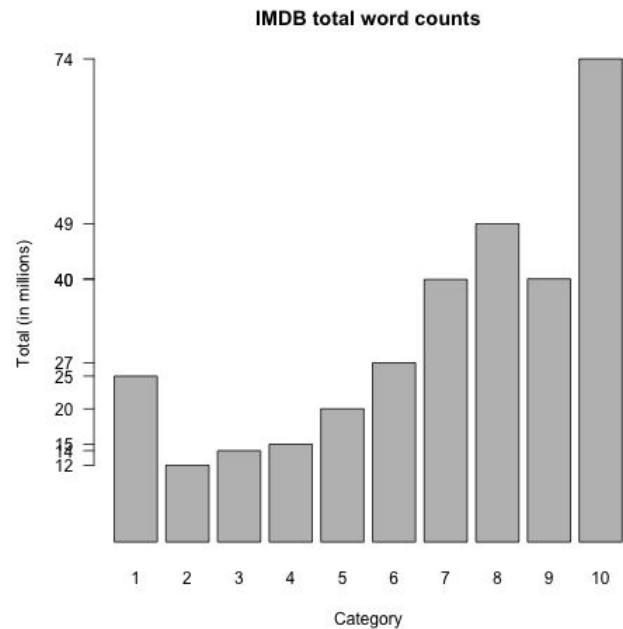
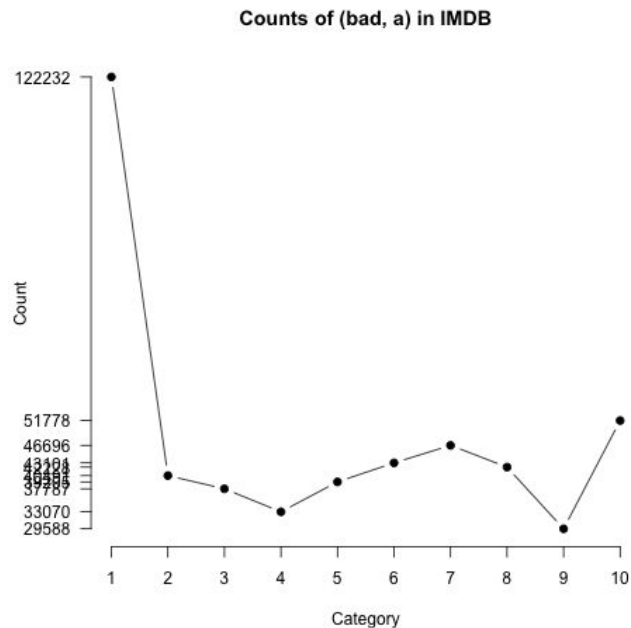


Las palabras y el sentimiento

¿Cómo se distribuye la palabra “BAD” en las críticas?



Las clases están desbalanceadas



Probabilidades

Rel Freq

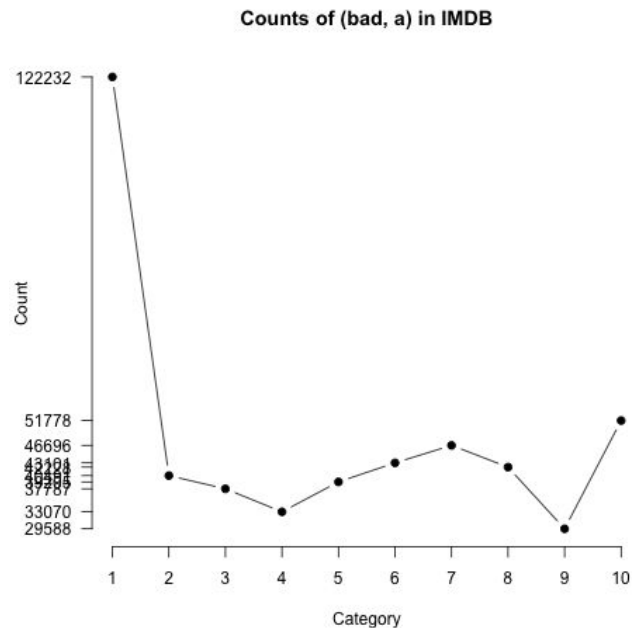
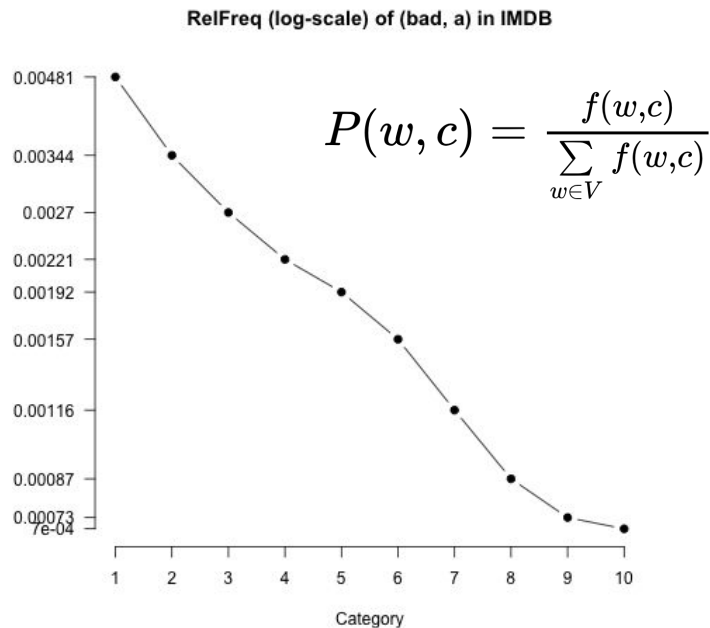
Counts

$$P(w|c) = \frac{f(w,c)}{\sum_{w \in V} f(w,c)}$$

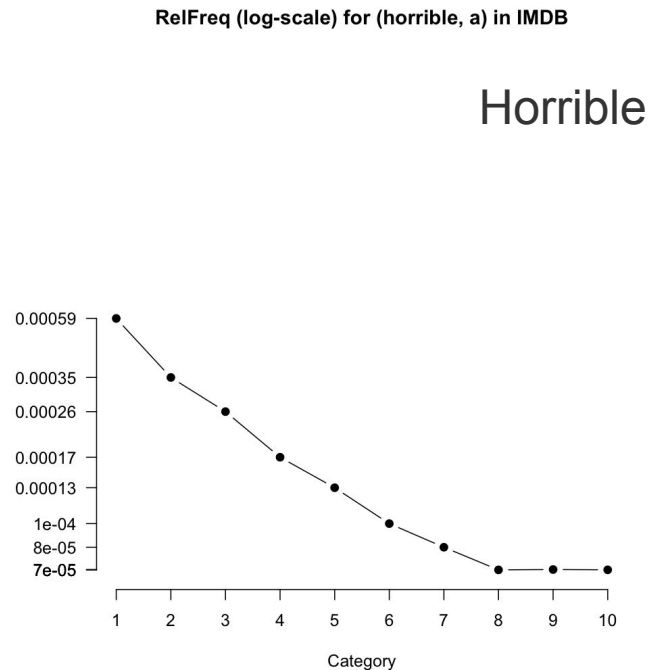
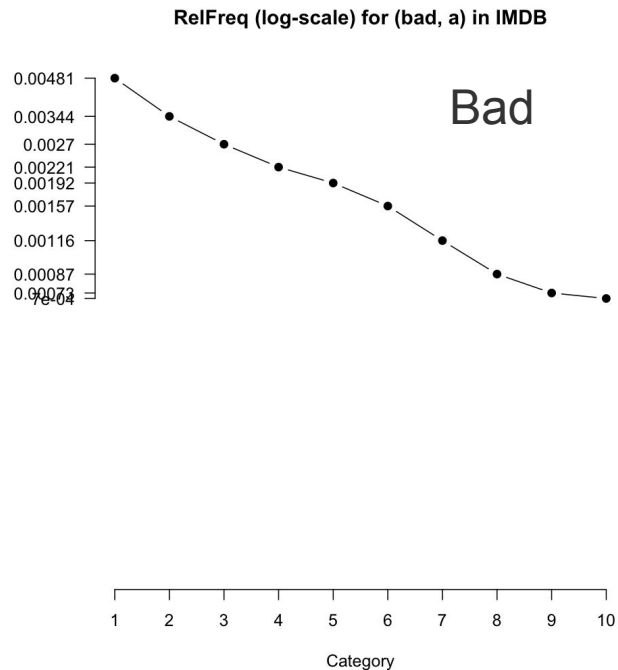
Total

Word	Tag	Category	Count	Total	RelFreq
bad	a	1	122232	25395214	0.0048
bad	a	2	40491	11755132	0.0034
bad	a	3	37787	13995838	0.0027
bad	a	4	33070	14963866	0.0022
bad	a	5	39205	20390515	0.0019
bad	a	6	43101	27420036	0.0016
bad	a	7	46696	40192077	0.0012
bad	a	8	42228	48723444	0.0009
bad	a	9	29588	40277743	0.0007
bad	a	10	51778	73948447	0.0007

Probabilidades



Comparación entre palabras



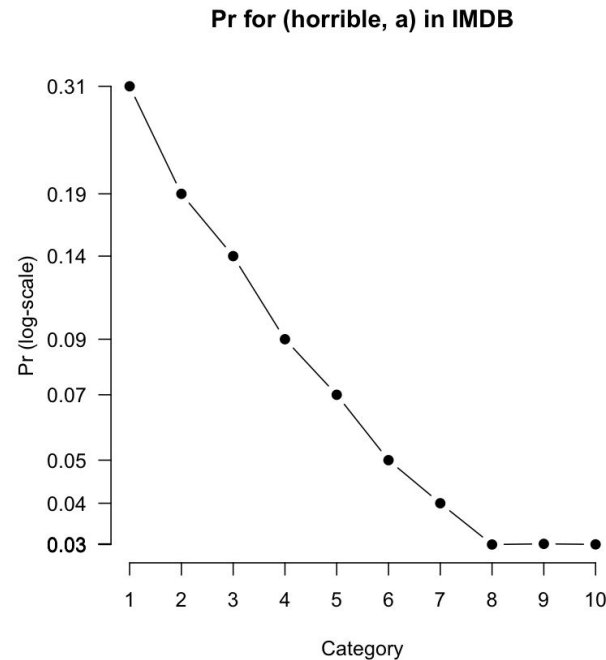
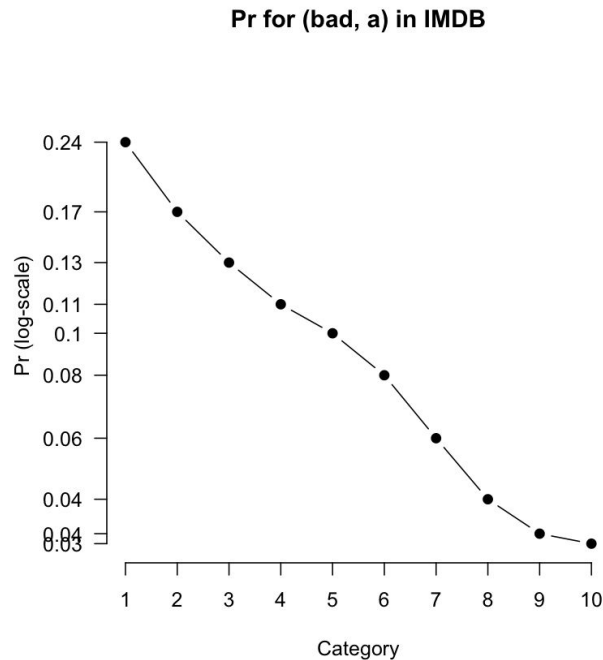
Probabilidades

$$P(w|c) = \frac{f(w,c)}{\sum_{w \in V} f(w,c)}$$

$$\text{Score}(w, c) = \frac{P(w|c)}{\sum_{C \in V} P(w|c)}$$

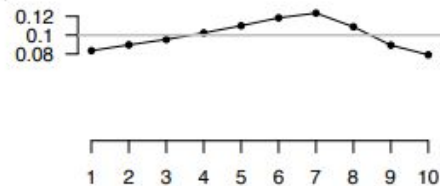
Word	Tag	Category	Count	Total	RelFreq
bad	a	1	122232	25395214	0.0048
bad	a	2	40491	11755132	0.0034
bad	a	3	37787	13995838	0.0027
bad	a	4	33070	14963866	0.0022
bad	a	5	39205	20390515	0.0019
bad	a	6	43101	27420036	0.0016
bad	a	7	46696	40192077	0.0012
bad	a	8	42228	48723444	0.0009
bad	a	9	29588	40277743	0.0007
bad	a	10	51778	73948447	0.0007

Comparación entre palabras

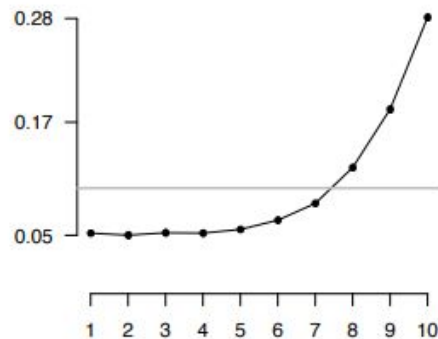


$$\text{Score}(w, c) = \frac{P(w|c)}{\sum_{C \in V} P(w|c)}$$

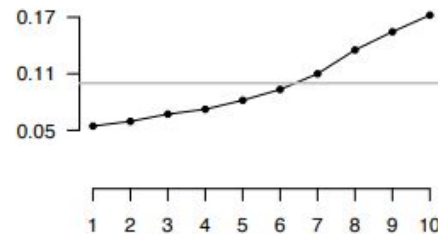
POS good (883,417 tokens)



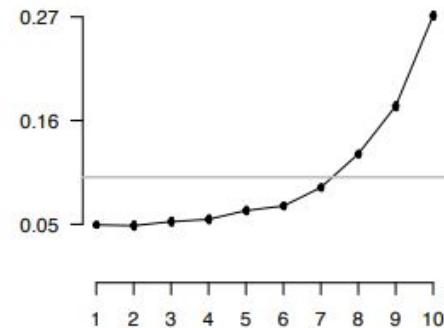
amazing (103,509 tokens)



great (648,110 tokens)

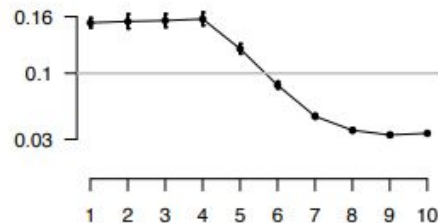


awesome (47,142 tokens)

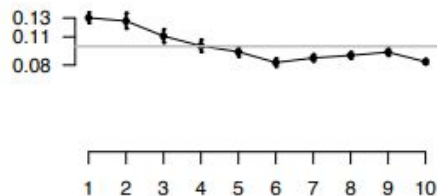


Rating

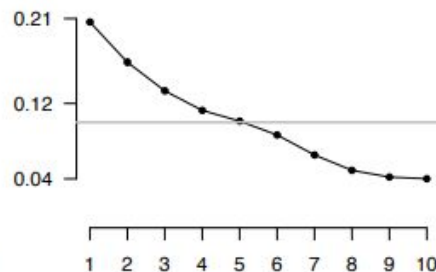
NEG good (20,447 tokens)



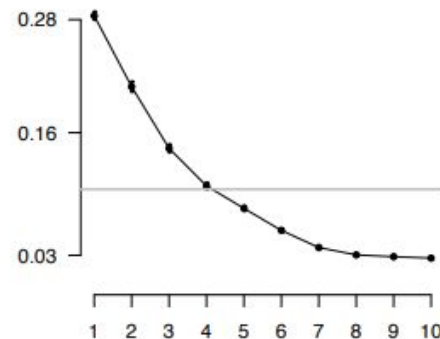
depress(ed/ing) (18,498 tokens)



bad (368,273 tokens)



terrible (55,492 tokens)



FIN