Análisis de Sentimiento en Twitter

Quantil
Emilio Silva
Agosto 2014

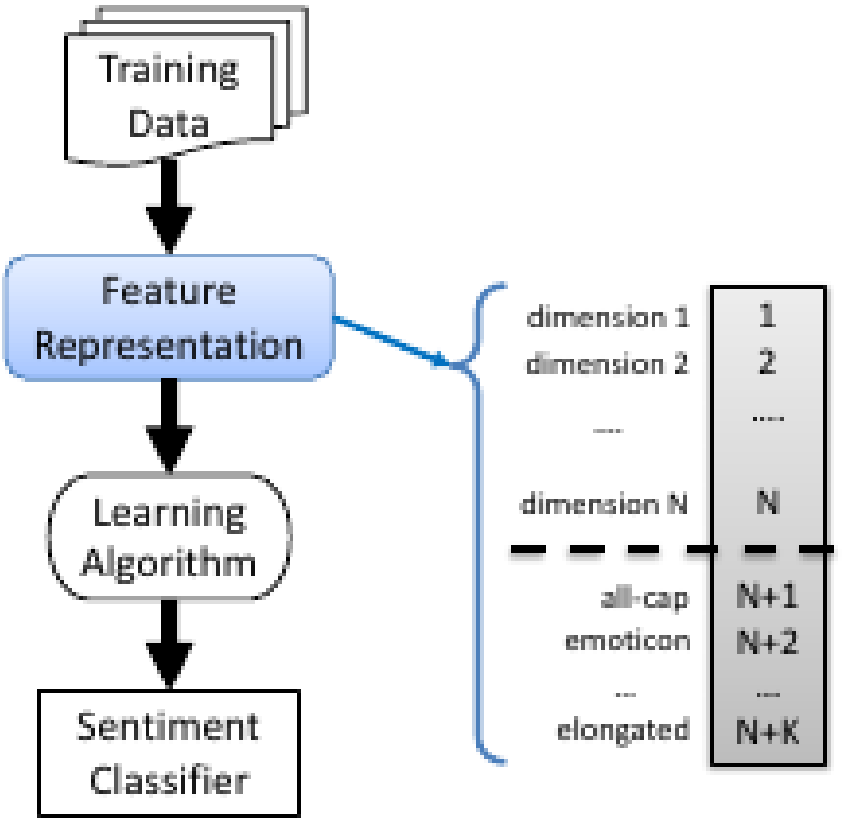
Introducción

- El análisis de sentimiento es la determinación de la polaridad o categoría de un mensaje en términos de su contenido emotivo
- Este contenido puede hacer referencia a un objeto de sentimiento, o a uno de sus aspectos
- Twitter es una de las principales fuentes de análisis
- Está evolucionando muy rápido y tiene muchas aplicaciones
- La gran mayoría de recursos está en inglés

Procedimiento General

1) Extracción de mensajes
2) Filtrado de elementos irrelevantes
3) Identificación de mensajes subjetivos
4) Identificación de la polaridad

Modelo de Clasificación

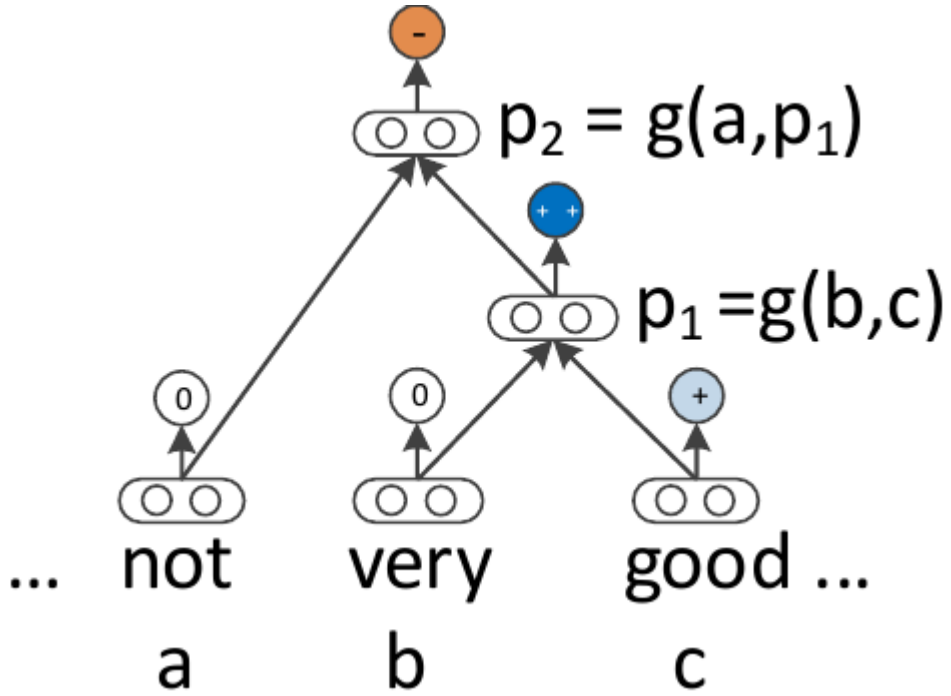


Stanford Sentiment

- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts
- Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank
- Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)

Stanford Sentiment

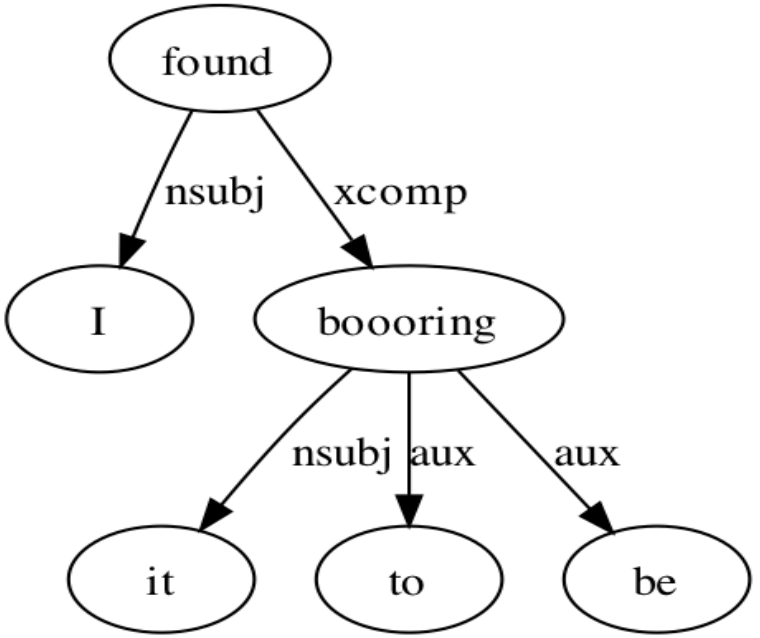
- Feature Representation: compositional vectors



Stanford Sentiment

Syntactic Dependencies Parser

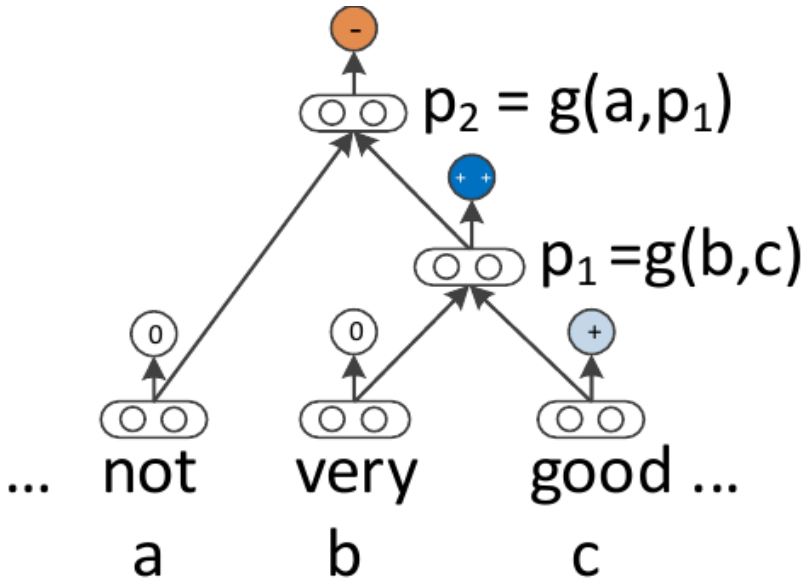
- I found it to be boooring !
- nsubj(found-2, I-1), nsubj(boooring-6, it-3), aux(boooring-6, to-4), aux(boooring-6, be-5), xcomp(found-2, boooring-6)



Stanford Sentiment

- Learning Algorithm: Recursive Neural Networks

$$p_1 = f\left(W\begin{bmatrix} b \\ c \end{bmatrix}\right), p_2 = f\left(W\begin{bmatrix} a \\ p_1 \end{bmatrix}\right)$$

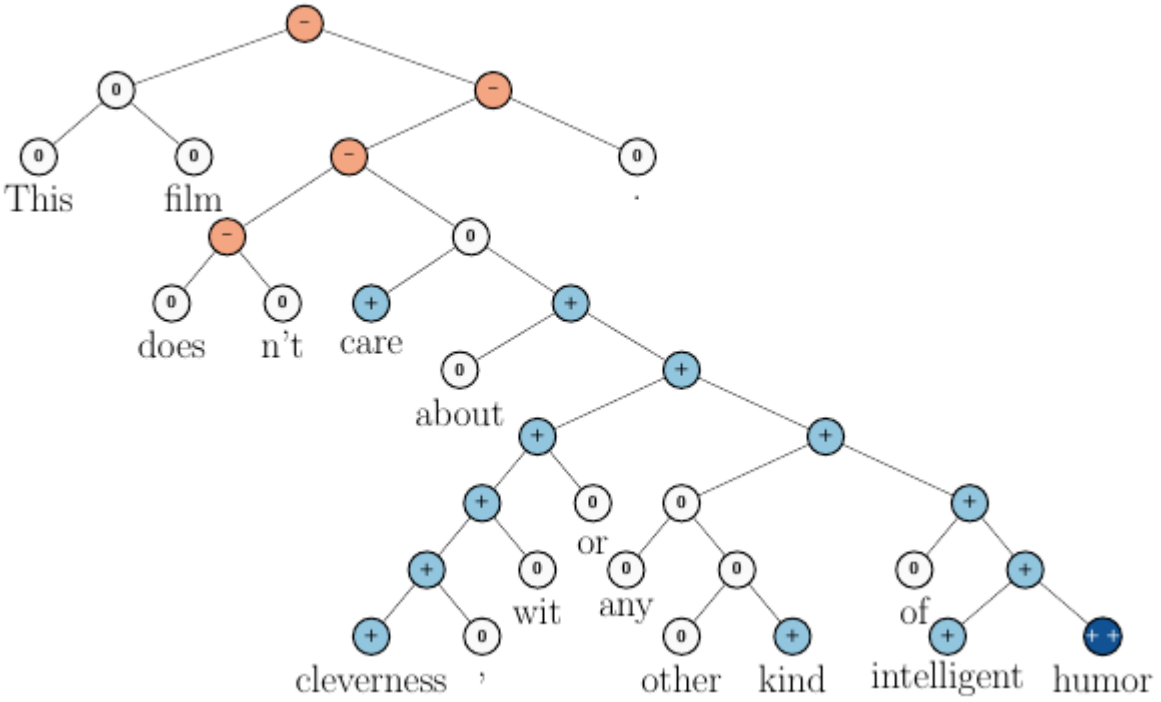


f = *tanh* is a standard element-wise nonlinearity

W ∈ **R**d×2d is the main parameter to learn

Stanford Sentiment

- Training Data: Sentiment Treebank



Stanford Sentiment

- Creación del Corpus: Torkers

Stanford Sentiment

Ventajas

- Estado del arte
- Código disponible

Desventajas

- Habría que hacer un Sentiment Treebank para español
- No está diseñado para Twitter

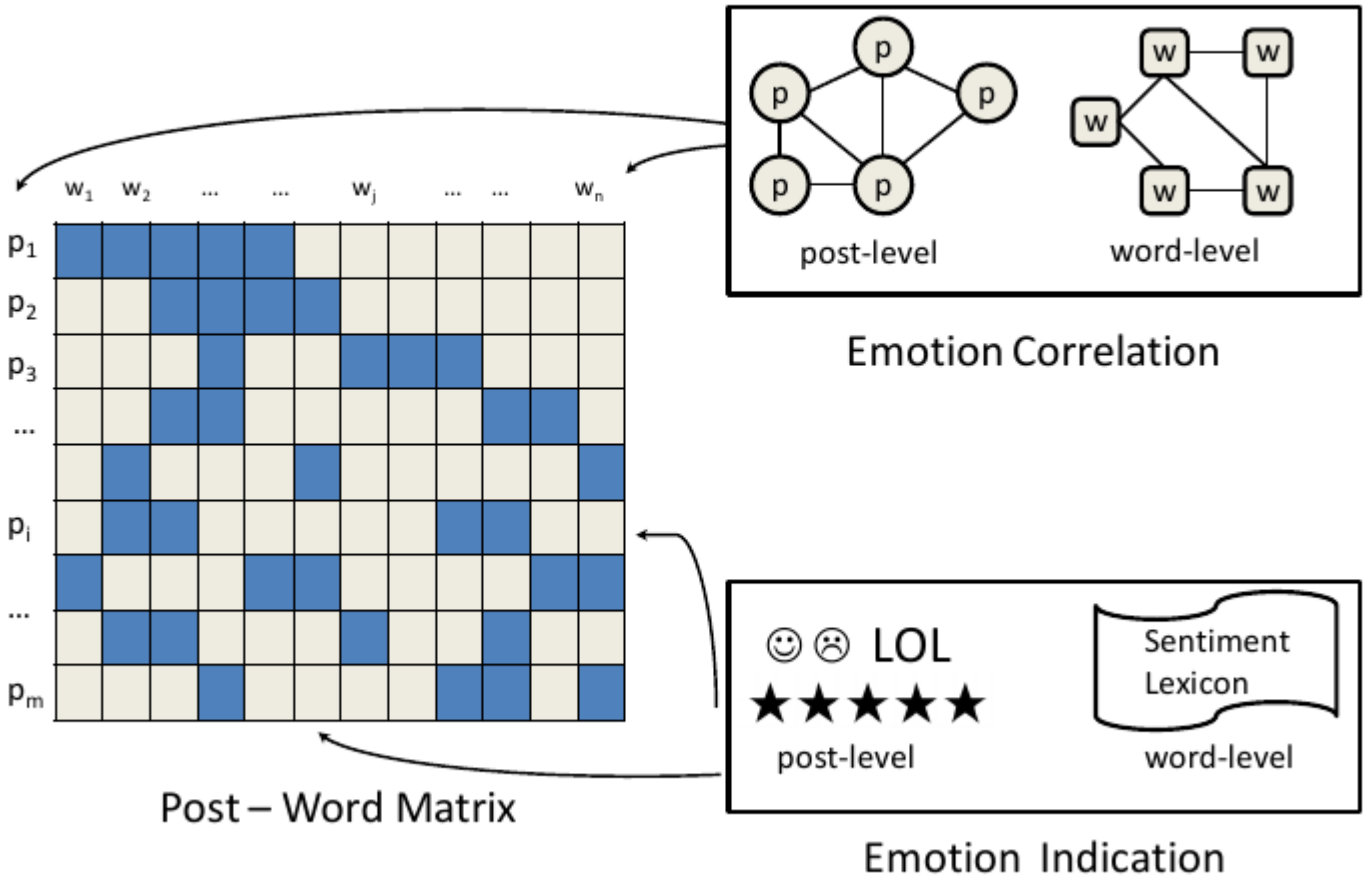
Características de Twitter

- Mensajes breves
- Poca estructura lingüística
- Lenguaje no convencional
- Presencia de elementos extralingüísticos

Hu et al. (2013)

- Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu
- Unsupervised Sentiment Analysis with Emotional Signals
- Proceedings of the International World Wide Web Conference, pages 607–618.

Hu et al. (2013)


p
w
post-level
word-level
Emotion Correlation
LOL
Sentiment Lexicon
post-level
word-level
Emotion Indication
Post – Word Matrix

Hu et al. (2013)

- **Emotion Indication**

  Señales emocionales de fácil acceso que reflejan el sentimiento de un elemento. A nivel de mensaje, v. gr. emotícono; a nivel de palabra, v. gr. léxico de emociones

Hu et al. (2013)

- **Emotion Indication**

  Se formula como la minimización de la función de pérdida

$$\|\mathbf{U}(i,*) - \mathbf{U}_0(i,*)\|_2^2,$$

  U: Rmxc es la matriz de sentimiento
  U0: Rmxc es la matriz de indicación de emoción

Hu et al. (2013)

- **Emotion Correlation**

  Señales emocionales que reflejan la similaridad de sentimiento entre dos elementos. A nivel de mensaje, v. gr. similaridad textual; a nivel de palabra, v. gr. coocurrencia en grandes corpus

Hu et al. (2013)

- **Emotion Correlation**

  Se formula como un grafo cuya matriz de adyacencia se define como

$$\mathbf{W}^u(i,j) = \begin{cases} 1 & \text{if } u_i \in \mathcal{N}(u_j) \text{ or } u_j \in \mathcal{N}(u_i) \\ 0 & \text{otherwise .} \end{cases}$$

- Ui es un elemento del grafo
- N(Uj) es el conjunto de los k vecinos más próximos de Uj

Hu et al. (2013)

- Se utiliza un modelo de tri-factorización de matrices ortogonales no negativas:

$$\min_{\mathbf{U},\mathbf{H},\mathbf{V}\geq 0} \quad \mathcal{O} = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2,$$

- X es la matriz mensaje – palabra
- U y V son matrices ortogonales no negativas
- U es la matriz mensaje – sentimiento
- V es la matriz palabra – sentimiento
- H $\in$ **R**cxc es una vista condensada de X

Hu et al. (2013)

- Solucionar el siguiente problema de optimización:

$$\min_{\mathbf{U},\mathbf{H},\mathbf{V}\geq 0} \quad \mathcal{J} = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2 + \lambda_I^u \|\mathbf{G}^u(\mathbf{U} - \mathbf{U}_0)\|_F^2$$
$$+ \lambda_I^v \|\mathbf{G}^v(\mathbf{V} - \mathbf{V}_0)\|_F^2 + \lambda_C^u Tr(\mathbf{U}^T \mathcal{L}^{\mathbf{u}} \mathbf{U}) + \lambda_C^v Tr(\mathbf{V}^T \mathcal{L}^{\mathbf{v}} \mathbf{V}),$$

Donde los λ son parámetros positivos de regularización.

Algoritmo de optimización

Input: $\{\mathbf{X}, \mathbf{U}_0, \mathbf{V}_0, \lambda_I^u, \lambda_I^v, \lambda_C^u, \lambda_C^v, T\}$
Output: $\mathbf{V}$

1: Construct matrices $\mathbf{G}^u$ and $\mathbf{G}^v$ in Eq. (3) and (4)
2: Construct matrices $\mathcal{L}^u$ and $\mathcal{L}^v$ in Eq. (7) and (9)
3: Initialize $\mathbf{U} = \mathbf{U}_0, \mathbf{V} = \mathbf{V}_0, \mathbf{H} \geq 0$
4: **while** Not convergent and t $\leq$ T **do**
5: Update $\mathbf{H}(i,j) \leftarrow \mathbf{H}(i,j)\sqrt{\frac{[\mathbf{U}^T\mathbf{X}\mathbf{V}](i,j)}{[\mathbf{U}^T\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}](i,j)}}$
6: Update

$$\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{[\mathbf{X}\mathbf{V}\mathbf{H}^T + \lambda_I^u\mathbf{G}^u\mathbf{U}_0 + \lambda_C^u\mathbf{W}^u\mathbf{U} + \mathbf{U}\Gamma_U^-](i,j)}{[\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \lambda_I^u\mathbf{G}^u\mathbf{U} + \lambda_C^u\mathbf{D}^u\mathbf{U} + \mathbf{U}\Gamma_U^+](i,j)}}$$

7: Update

$$\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j)\sqrt{\frac{[\mathbf{X}^T\mathbf{U}\mathbf{H} + \lambda_I^v\mathbf{G}^v\mathbf{V}_0 + \lambda_C^v\mathbf{W}^v\mathbf{V} + \mathbf{V}\Gamma_V^-](i,j)}{[\mathbf{V}\mathbf{H}^T\mathbf{U}^T\mathbf{U}\mathbf{H} + \lambda_I^v\mathbf{G}^v\mathbf{V} + \lambda_C^v\mathbf{D}^v\mathbf{V} + \mathbf{V}\Gamma_V^+](i,j)}}$$

8: $t = t + 1$
9: **end while**

Conclusión

Para implementar el método de Hu et al. (2013) se necesita:

- Un lexicón de emotíconos clasificados
- Un lexicón de palabras con polaridad
- Un método de similaridad semántica
- Un corpus grande para encontrar coocurrencias

- Todo esto se puede conseguir