Una introducción al Análisis de Sentimientos en microblogs

Victor Muñiz Sánchez

Centro de Investigación en Matemáticas.
Unidad Monterrey.

EPE 2015.

Contenido

Introducción

Introducción

Críticas de películas en `http://www.rottentomatoes.com/`

ONLY LOVERS LEFT ALIVE (2014)

TOMATOMETER

All Critics | **Top Critics**

CERTIFIED FRESH Rotten Tomatoes **90%**

Average Rating: 7.5/10
Reviews Counted: 42
Fresh: 38
Rotten: 4

Critics Consensus: Worth watching for Tom Hiddleston and Tilda Swinton's performances alone, Only Lovers Left Alive finds writer-director Jim Jarmusch adding a typically offbeat entry to the vampire genre.

Liam Lacey
Globe and Mail
★ Top Critic

Only Lovers is so fluidly edited and thinly plotted that it feels almost off-hand; yet, it's also made with great care, beautifully lit and set-designed to an eyelash. April 25, 2014

Full Review | Original Score: 3.5/4

Stephanie Zacharek
Village Voice
★ Top Critic

This is Jarmusch's most emotionally direct film since Dead Man, and maybe his finest. April 8, 2014

Full Review

Mick LaSalle
San Francisco Chronicle
★ Top Critic

This is a weak entry in the filmmaker's canon, a film that continues Jarmusch's rebellion against narrative while offering nothing in its place. April 18, 2014

Full Review | Original Score: 1/4

Introducción

En análisis de sentimientos en textos, seguimos los siguientes pasos:

- Recopilamos información en una representación adecuada
- Extraemos información relevante de cada texto (¿Qué es **relevante**? ¿Cómo la definimos y medimos?)
- En base a ésta información, le asignamos una categoría (Buena, Mala, Neutral, por ejemplo)
- Todo se hace de manera automática

Introducción

En análisis de sentimientos en textos, seguimos los siguientes pasos:

- Recopilamos información en una representación adecuada
- Extraemos información relevante de cada texto (¿Qué es **relevante**? ¿Cómo la definimos y medimos?)
- En base a ésta información, le asignamos una categoría (Buena, Mala, Neutral, por ejemplo)
- Todo se hace de manera automática

Introducción

En análisis de sentimientos en textos, seguimos los siguientes pasos:

- Recopilamos información en una representación adecuada
- Extraemos información relevante de cada texto (¿Qué es **relevante**? ¿Cómo la definimos y medimos?)
- En base a ésta información, le asignamos una categoría (Buena, Mala, Neutral, por ejemplo)
- Todo se hace de manera automática

Introducción

En análisis de sentimientos en textos, seguimos los siguientes pasos:

- Recopilamos información en una representación adecuada
- Extraemos información relevante de cada texto (¿Qué es **relevante**? ¿Cómo la definimos y medimos?)
- En base a ésta información, le asignamos una categoría (Buena, Mala, Neutral, por ejemplo)
- Todo se hace de manera automática

Introducción

El área de análisis de sentimientos (opinion mining) ha tenido mucho auge en los últimos años, pero tiene sus raíces en trabajos sobre **recuperación de la información (information retrieval)** y **procesamiento del lenguaje natural**:

- Jaime Carbonell. **Subjective Understanding: Computer Models of Belief Systems**. PhD thesis, Yale, 1979.
- Yorick Wilks and Janusz Bien. **Beliefs, points of view and multiple environments**. In Proceedings of the international NATO symposium on artificial and human intelligence, 1984.

Introducción

El área de análisis de sentimientos (opinion mining) ha tenido mucho auge en los últimos años, pero tiene sus raíces en trabajos sobre **recuperación de la información (information retrieval)** y **procesamiento del lenguaje natural**:

- Jaime Carbonell. **Subjective Understanding: Computer Models of Belief Systems**. PhD thesis, Yale, 1979.
- Yorick Wilks and Janusz Bien. **Beliefs, points of view and multiple environments**. In Proceedings of the international NATO symposium on artificial and human intelligence, 1984.

Introducción

El área de análisis de sentimientos (opinion mining) ha tenido mucho auge en los últimos años, pero tiene sus raíces en trabajos sobre **recuperación de la información (information retrieval)** y **procesamiento del lenguaje natural**:

- Jaime Carbonell. **Subjective Understanding: Computer Models of Belief Systems**. PhD thesis, Yale, 1979.
- Yorick Wilks and Janusz Bien. **Beliefs, points of view and multiple environments**. In Proceedings of the international NATO symposium on artificial and human intelligence, 1984.

Introducción

- Marti Hearst. **Direction-based text interpretation as an information access refinement**. Text-Based Intelligent Systems, 1992.
- Janyce M. Wiebe and William J. Rapaport. **A computational theory of perspective and reference in narrative**. In Proceedings of the Association for Computational Linguistics (ACL), 1988.
- Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. **Development and use of a gold standard data set for subjectivity classifications**. In Proceedings of the Association for Computational Linguistics (ACL), 1999.

Introducción

- Marti Hearst. **Direction-based text interpretation as an information access refinement**. Text-Based Intelligent Systems, 1992.
- Janyce M. Wiebe and William J. Rapaport. **A computational theory of perspective and reference in narrative**. In Proceedings of the Association for Computational Linguistics (ACL), 1988.
- Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. **Development and use of a gold standard data set for subjectivity classifications**. In Proceedings of the Association for Computational Linguistics (ACL), 1999.

Introducción

- Marti Hearst. **Direction-based text interpretation as an information access refinement**. Text-Based Intelligent Systems, 1992.
- Janyce M. Wiebe and William J. Rapaport. **A computational theory of perspective and reference in narrative**. In Proceedings of the Association for Computational Linguistics (ACL), 1988.
- Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. **Development and use of a gold standard data set for subjectivity classifications**. In Proceedings of the Association for Computational Linguistics (ACL), 1999.

Introducción

El auge de ésta área de investigación tiene varias razones. Una de las principales es la diversidad de plataformas virtuales de opinión y la gran cantidad de información que se genera.

Introducción

Y también la gran variedad de aplicaciones que tiene

- Business intelligence: estrategias basadas en opinión de clientes
  - **Online consumer-generated reviews have significant impact on offline purchase behavior**. Press Release, November 2007. `http://www.comscore.com/press/release.asp?press=1928`
  - Gilad Mishne and Natalie Glance. **Predicting movie sales from blogger sentiment**. In AAAI Symposium on Computational Approaches to Analysing Weblogs, 2006
  - Ana-Maria Popescu and Oren Etzioni. **Extracting product features and opinions from reviews**. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005

Introducción

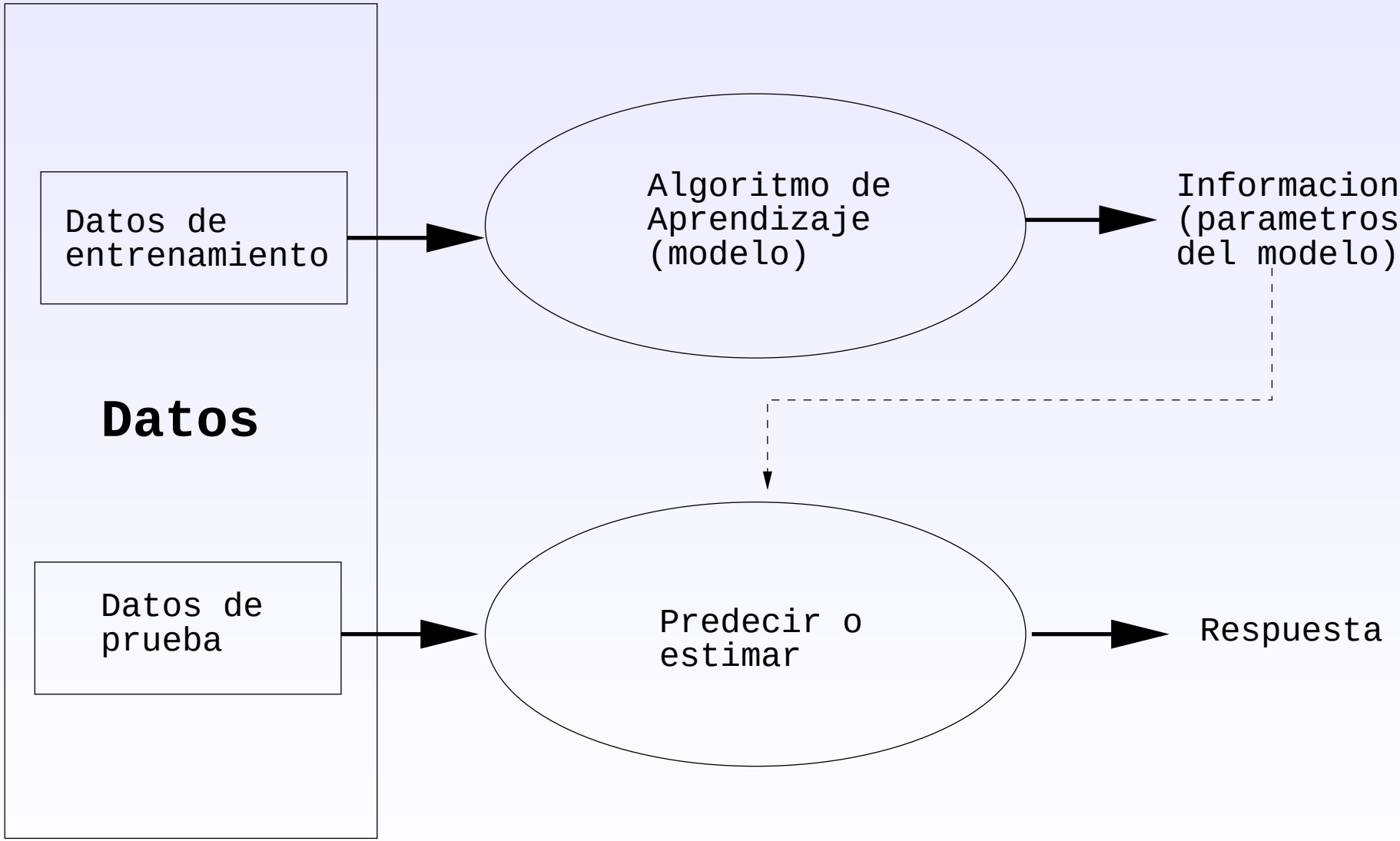
- Sitios web de revisión y calificación de productos y servicios: eBay, Amazon, hotels.com, y un largo etcétera
  - Jingjing Liu, Yunbo Cao. **Low-quality product review detection in opinion summarization**. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007.
  - Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. **Automatically assessing review helpfulness**. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2006.
  - Luis Cabral and Ali Hortacsu. **The Dynamics of Seller Reputation: Theory and Evidence from eBay**. The National Bureau of Economic Research. Working Paper No. 10363, 2006.

Introducción

- Como un componente tecnológico para otros sistemas:
  - Li Zhuang, Feng Jing, Xiao-yan Zhu, and Lei Zhang. **Movie review mining and summarization**. In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM),2006.
  - Xin Jin. **Sensitive webpage classification for content advertising** In Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising, 2007.
  - Scott Piao, Sophia Ananiadou, Yoshimasa Tsuruoka, Yutaka Sasaki, and John McNaught. **Mining opinion polarity relations of citations**. In International Workshop on Computational Semantics (IWCS), 2007.
  - Maite Taboada, Mary Ann Gillies, and Paul McFetridge. **Sentiment classification techniques for tracking literary reputation** In LREC Workshop: Towards Computational Models of Literary Analysis, 2006.
- En gobierno y política: votaciones, índices de satisfacción, termómetro político.

Análisis de sentimientos como un problema de Machine Learning

AS y machine learning



AS y machine learning

Nuestros datos son **datos no estructurados** (textos). ¿Cómo podemos encontrar medidas de similaridad (relevantes) para compararlos entre ellos?

- *"Only lovers is so fluidly edited and thinly plotted that it feels almost off-hand; yet, it's also made with care, beautifully lit and set-designed to an eyelash"*
- *"This is Jarmusch's most emotionally direct film since Dead Man, and maybe his finest"*
- *"This is a weak entry in the filmmaker's canon, a film that continues Jarmusch's rebellion against narrative while offering nothing in its place"*

AS y machine learning

Nuestros datos son **datos no estructurados** (textos). ¿Cómo podemos encontrar medidas de similaridad (relevantes) para compararlos entre ellos?

- *"Only lovers is so fluidly edited and thinly plotted that it feels almost off-hand; yet, it's also made with care, beautifully lit and set-designed to an eyelash"*
- *"This is Jarmusch's most emotionally direct film since Dead Man, and maybe his finest"*
- *"This is a weak entry in the filmmaker's canon, a film that continues Jarmusch's rebellion against narrative while offering nothing in its place"*

AS y machine learning

Nuestros datos son **datos no estructurados** (textos). ¿Cómo podemos encontrar medidas de similaridad (relevantes) para compararlos entre ellos?

- *"Only lovers is so fluidly edited and thinly plotted that it feels almost off-hand; yet, it's also made with care, beautifully lit and set-designed to an eyelash"*
- *"This is Jarmusch's most emotionally direct film since Dead Man, and maybe his finest"*
- *"This is a weak entry in the filmmaker's canon, a film that continues Jarmusch's rebellion against narrative while offering nothing in its place"*

AS y machine learning

Una opción lógica es definir una lista de palabras "positivas" y otra "negativas"

- $(+)$: dazzling, brilliant, phenomenal, excellent, fantastic, gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting
- $(-)$: suck, terrible, awful, unwatchable, hideous, bad, cliched, sucks, boring, stupid, slow, worst, stupid, waste

Entonces, un clasificador simple es, contar las palabras positivas y negativas que contiene un texto.
Una crítica será positiva si $\sum(+) > \sum(-)$.

¿Cómo elegir la lista de palabras "relevantes" para cada categoría?

AS y machine learning

Una opción lógica es definir una lista de palabras "positivas" y otra "negativas"

- $(+)$: dazzling, brilliant, phenomenal, excellent, fantastic, gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting
- $(-)$: suck, terrible, awful, unwatchable, hideous, bad, cliched, sucks, boring, stupid, slow, worst, stupid, waste

Entonces, un clasificador simple es, contar las palabras positivas y negativas que contiene un texto.

Una crítica será positiva si $\sum(+) > \sum(-)$.

¿Cómo elegir la lista de palabras "relevantes" para cada categoría?

AS y machine learning

Una opción lógica es definir una lista de palabras "positivas" y otra "negativas"

- $(+)$: dazzling, brilliant, phenomenal, excellent, fantastic, gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting
- $(-)$: suck, terrible, awful, unwatchable, hideous, bad, cliched, sucks, boring, stupid, slow, worst, stupid, waste

Entonces, un clasificador simple es, contar las palabras positivas y negativas que contiene un texto.
Una crítica será positiva si $\sum(+) > \sum(-)$.

¿Cómo elegir la lista de palabras "relevantes" para cada categoría?

AS y machine learning

Una opción lógica es definir una lista de palabras "positivas" y otra "negativas"

- $(+)$: dazzling, brilliant, phenomenal, excellent, fantastic, gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting
- $(-)$: suck, terrible, awful, unwatchable, hideous, bad, cliched, sucks, boring, stupid, slow, worst, stupid, waste

Entonces, un clasificador simple es, contar las palabras positivas y negativas que contiene un texto.
Una crítica será positiva si $\sum(+) > \sum(-)$.

¿Cómo elegir la lista de palabras "relevantes" para cada categoría?

AS y machine learning

Las palabras relevantes para las categorías de interés dependen del contexto, entonces, es mejor que los mismos datos nos digan qué palabras son relevantes. La medida más común es el conteo de palabras en textos.

Definimos:

- d: Un documento
- Corpus: el conjunto de documentos a procesar.
- t_i: Una palabra (término) de un documento
- Diccionario: Un conjunto predefinido de N términos que pertenecen a los documentos.

AS y machine learning

Las palabras relevantes para las categorías de interés dependen del contexto, entonces, es mejor que los mismos datos nos digan qué palabras son relevantes. La medida más común es el conteo de palabras en textos. Definimos:

- d: Un documento
- Corpus: el conjunto de documentos a procesar.
- t_i: Una palabra (término) de un documento
- Diccionario: Un conjunto predefinido de N términos que pertenecen a los documentos.

AS y machine learning

- Matriz de términos de $m \times N$

$$D = \begin{pmatrix} tf(t_1, d_i) & \cdots & tf(t_N, d_1) \\ \vdots & \ddots & \vdots \\ tf(t_1, d_m) & \cdots & tf(t_N, d_m) \end{pmatrix}$$

donde $tf(t_j, d)$ es la frecuencia del término t_j en el documento d.

Con lo anterior, obtenemos que

$$\text{documento de texto } \in \mathcal{R}^N,$$

y es llamado modelo *Bag of Words* (Joachims, 1998).

AS y machine learning

- Matriz de términos de $m \times N$

$$D = \begin{pmatrix} tf(t_1, d_i) & \cdots & tf(t_N, d_1) \\ \vdots & \ddots & \vdots \\ tf(t_1, d_m) & \cdots & tf(t_N, d_m) \end{pmatrix}$$

donde $tf(t_j, d)$ es la frecuencia del término t_j en el documento d.

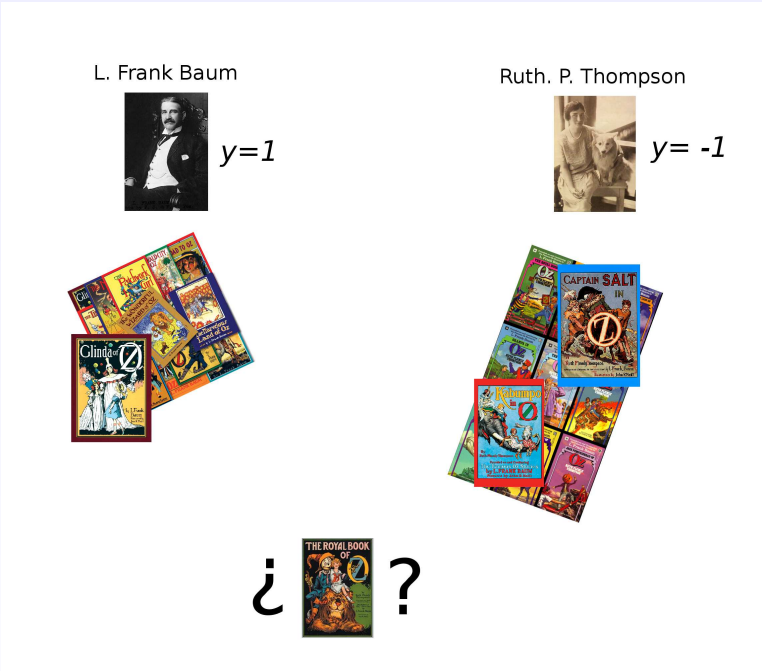
Con lo anterior, obtenemos que

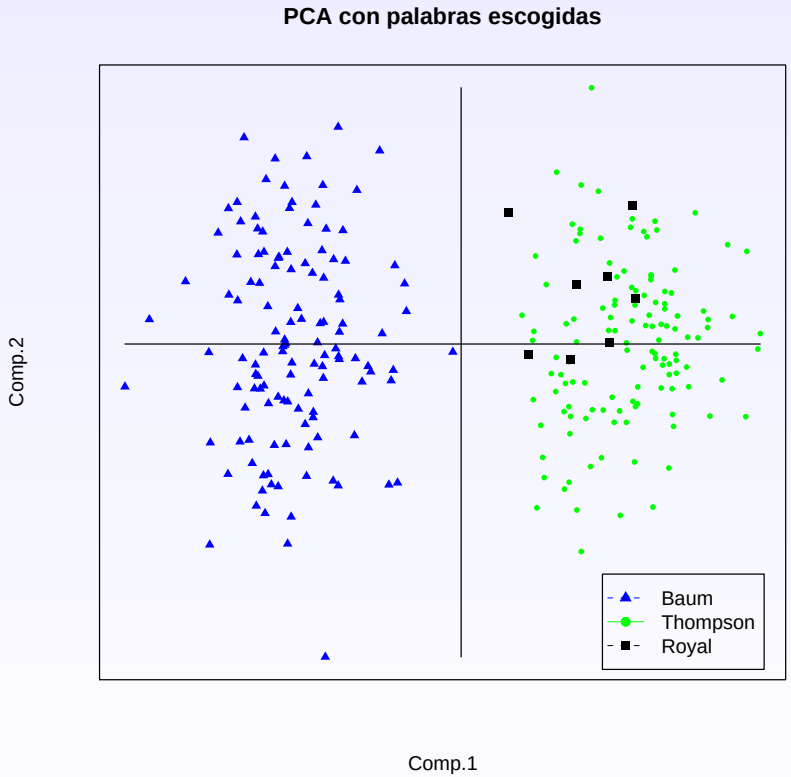
$$\text{documento de texto } \in \mathcal{R}^N,$$

y es llamado modelo *Bag of Words* (Joachims, 1998).

AS y machine learning

Un ejemplo muy conocido en atribución de autoría en literatura (Binongo, 2003):





AS en microblogs (twitter)

AS en twitter

El caso de los microblogs como twitter, representan retos adicionales por las características de su escritura:

- Son cortos, máximo 140 caracteres
- Escritura informal

```
Andas bien loco @Telcel con la zona horaria d tu RED,
a cada rato m mueves la Hr.?? #chidotucotorreo
@ServicioTelcelhttp://t.co/QoOX3OCYxt

@Profeco  @Tiendas_OXXO no cumple con algunos
requerimientos como tipos de bebida falsos as
como la falta del precio :(
```

AS en twitter

- Abreviaciones y contracciones no estándar
- Faltas ortográficas
- Clases desbalanceadas. Generalmente hay muchos tuits neutrales, menos negativos y muy pocos positivos.

AS en twitter

Un proyecto de análisis de sentimientos en twitter es interesante por varias razones

- Su potencial uso comercial
- La disponibilidad de información
- El reto que representa
- No hay mucho trabajo de investigación en español

AS en twitter

Un proyecto de análisis de sentimientos en twitter es interesante por varias razones

- Su potencial uso comercial
- La disponibilidad de información
- El reto que representa
- No hay mucho trabajo de investigación en español

AS en twitter

Un proyecto de análisis de sentimientos en twitter es interesante por varias razones

- Su potencial uso comercial
- La disponibilidad de información
- El reto que representa
- No hay mucho trabajo de investigación en español

AS en twitter

Un proyecto de análisis de sentimientos en twitter es interesante por varias razones

- Su potencial uso comercial
- La disponibilidad de información
- El reto que representa
- No hay mucho trabajo de investigación en español

AS en twitter

Un proyecto de análisis de sentimientos en twitter es interesante por varias razones

- Su potencial uso comercial
- La disponibilidad de información
- El reto que representa
- No hay mucho trabajo de investigación en español

Recopilación de datos

AS en twitter

Se recopilaron tuits usando la API de twitter (`https://dev.twitter.com/`) con diferentes características

- Por tópico
- Por periodos de tiempo
- Por ciudad y área geográfica

Se clasificaron aproximádamente 1500 tuits (para cada tópico) manualmente en tres categorías: Positivo, Negativo y Otro (neutral).

Preproceso

AS en twitter

Siguiendo procedimientos sugeridos en la literatura, los tuits se preprocesaron de la siguiente forma:

- Se removieron mensajes iguales
- Se puso el texto en minúsculas
- Se eliminaron palabras comúnes (stopwords) en español según el listado del *Snowball stemming project* (Martin Porter) `http://snowball.tartarus.org/`. A éstas les agregamos otras relativas al tópico.
- Se eliminaron caracteres especiales: URL's, @, RT, _, -, :, entre otros

AS en twitter

- Se eliminaron exceso de espacios en blanco y letras repetidas (`hooollaaaaaa=hola`)
- Se reemplazaron los emoticons con un texto descriptivo según la lista: `en.wikipedia.org/wiki/List_of_emoticons`. Por ejemplo:

:-) emoticon-positivo	>:[emoticon-negativo
:) emoticon-positivo	=(emoticon-negativo
:o) emoticon-positivo	:-[emoticon-negativo
:c) emoticon-positivo	:-|| emoticon-muy-negativo
:-D emoticon-muy-positivo	>:(emoticon-muy-negativo
X-D emoticon-muy-positivo	:| emoticon-neutral

AS en twitter

- También, se realizó un proceso de normalización, para tratar de detectar los términos no convencionales y "traducirlos" en texto convencional. Muchos trabajos en la literatura (Melville, 2013. Mosquera, 2013 por ejemplo) han hecho notar que los métodos de clasificación mejoran considerablemente introduciendo una autocorrección en los textos.
  Esta normalización o auto-corrección se realizó en dos pasos:
  - Detección de palabras no convencionales
  - Sustitución de éstas palabras con aquellas "parecidas"

AS en twitter

Detección de palabras no convencionales

- Se utilizó el diccionario y la API en C de Aspell (`http://aspell.net/`) con un diccionario en español al que se le añadieron palabras como ciudades y localidades, y otras relativas al contexto.
- Se separa cada tuit en palabras, y se realiza una búsqueda para cada palabra. Si no aparece en el diccionario, se consideran las opciones dadas por Aspell.

AS en twitter

Detección de palabras no convencionales

- Inicialmente, se sustituyó la palabra incorrecta por la primera sugerencia de Aspell, pero no siempre es buena opción. Por ejemplo, para `pseudoestudiantes`

```
 [1] "pseudo"          "estudiantes"       "pseudo-estudiantes"
 [4] "predestinares"   "predestines"       "predestinases"
 [7] "predestinareis"  "predestinase"      "predestinar"
[10] "predestinas"     "predestinasteis"   "predestinaste"
[13] "predestinis"     "sudestada"         "predestinaras"
[16] "predestinars"    "predestinaseis"    "sudestadas"
[19] "predestinis"     "predestinadas"     "predestinados"
[22] "predestinabas"   "predestinamos"
```

AS en twitter

Detección de palabras no convencionales

- Una opción que exploramos es el uso de métodos de kernel con "string kernel".
- Sean $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, donde $\mathcal{X}$ denota el espacio de entrada. Un kernel k es una función que calcula el producto punto de dos puntos transformados mediante cierta función ϕ:

$$k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle \tag{1}$$

donde ϕ es un mapeo de $\mathcal{X}$ a un espacio de productos punto $\mathcal{H}$, al que llamaremos *espacio de características*:

$$\phi : \mathbf{x} \in \mathcal{X} \mapsto \phi(\mathbf{x}) \in \mathcal{H}.$$

AS en twitter

Detección de palabras no convencionales

- Una opción que exploramos es el uso de métodos de kernel con "string kernel".
- Sean $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, donde $\mathcal{X}$ denota el espacio de entrada. Un kernel k es una función que calcula el producto punto de dos puntos transformados mediante cierta función ϕ:

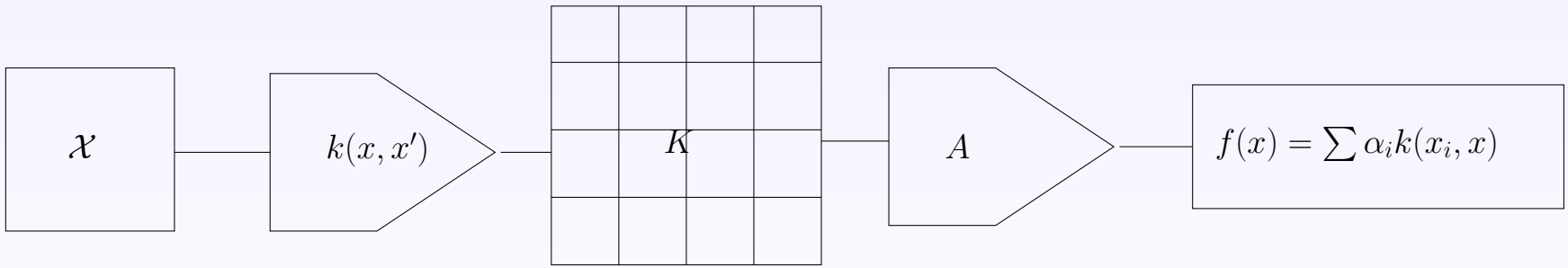
$$k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle \tag{1}$$

donde ϕ es un mapeo de $\mathcal{X}$ a un espacio de productos punto $\mathcal{H}$, al que llamaremos *espacio de características*:

$$\phi : \mathbf{x} \in \mathcal{X} \mapsto \phi(\mathbf{x}) \in \mathcal{H}.$$

AS en twitter

- Truco del kernel (Scholkopf y Smola, 2002): Dado un algoritmo formulado en términos de productos punto, puede construirse un algoritmo alternativo reemplazando los productos punto por un kernel positivo definido k.
- Esquema de Métodos de Kernel



AS en twitter

- Los String Kernels (Lodhi 2002, Shawe-Taylor y Cristianini 2004, Watkins 2000, Herbrich 2002) son una medida de similaridad entre dos secuencias de caracteres (documentos) x y y. Sea s una subcadena de texto, el mapeo al espacio de características está dado por

$$\phi(x) = \sum_{s \in x} \lambda(s),$$

  donde $\lambda(s) \in (0, 1)$ es un peso que depende de la longitud y ocurrencia de s.
- El kernel está dado por

$$k(x, y) = \sum_{s \in \Sigma} \text{num}_s(x)\text{num}_s(y)\lambda(u)$$

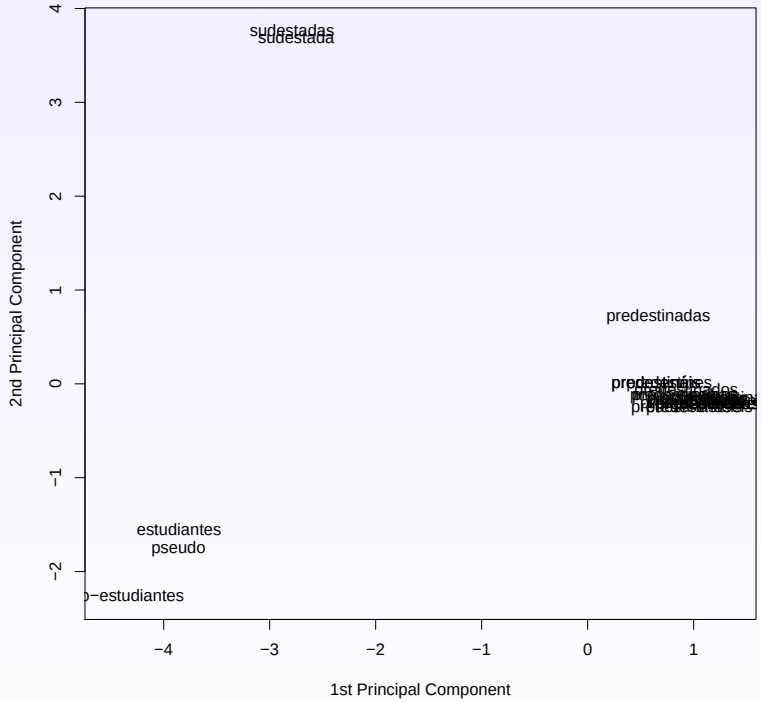
AS en twitter

- Poniendo $\lambda(s) = 0$ para subcadenas que empiecen y terminen con un espacio en blanco, nos da el "bag of words" kernel.
- Calcular la matriz de Gram para este kernel es en general muy costoso, pero hay métodos efectivos para hacerlo (suffix tree). Aún así, puede ser costoso para textos muy grandes.
- Ejemplo para las palabras cat, car, bat y bar con $|s| = 2$:

	c-a	c-t	a-t	b-a	b-t	c-r	a-r	b-r
ϕ(cat)	λ^2	λ^3	λ^2	0	0	0	0	0
ϕ(car)	λ^2	0	0	0	0	λ^3	λ^2	0
ϕ(bat)	0	0	λ^2	λ^2	λ^3	0	0	0
ϕ(bar)	0	0	0	λ^2	0	0	λ^2	λ^3

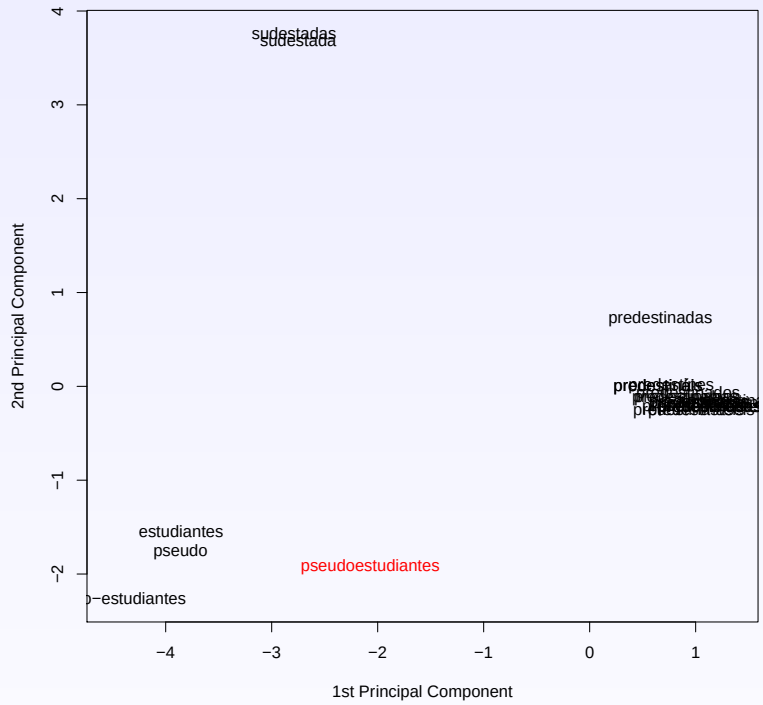
AS en twitter

- Basado en las palabras sugeridas por Aspell, obtuvimos una representación vectorial con Kernel PCA. Por ejemplo, considera la palabra `presudoestudiantes`, y las sugerencias de Aspell:



AS en twitter

- La proyección de `presudoestudiantes` en los dos primeros componentes principales dados por kernel PCA es



- Usando el criterio de distancia mínima con 3 componentes principales nos da que la palabra más parecida es `pseudo-estudiantes`, que es mejor que `pseudo` o `estudiantes` por separado.

AS en twitter

- En general, funciona bien éste procedimiento, pero a veces da resultados incorrectos comparados con Aspell:

```
malinxhista : molinista (=)
xhinga : chinga (=)
osico : musico (=)
m\'exico : l\'exico (x)
corido : corrido (+)
htm\'exico : M\'exico (+)
```

Clasificación

AS en twitter

Con los tuits preprocesados, se implementó un método de clasificación como el usado por Melville et al. 2013, con las siguientes características.

- Se utilizan las palabras más relevantes (bag of words) en tuits positivos,negativos y neutros, usando una medida de información mútua (Yang and Pedersen, 1997).
- Se añaden también, como información apriori, una lista de palabras predefinidas para cada categoria, así como información de emoticons y hashtags.

AS en twitter

Con los tuits preprocesados, se implementó un método de clasificación como el usado por Melville et al. 2013, con las siguientes características.

- Se implementa un método de clasificación ingénuo Bayesiano multinomial que asigna la clase c a un tuit d mediante

$$c^* = argmax_c p(c|d),$$

donde

$$p(c|d) = \alpha_1 p_1(c|d) + \alpha_2 p_2(c|d),$$

$p_1(c|d)$ son las probabilidades de cada clase usando bag of words y $p_2(c|d)$ son las probabilidades obtenidas usando las palabras predefinidas. α_1 y α_2 son sus respectivos pesos. Las probabilidades están dadas por el clasificador ingénuo bayesiano:

$$p(c|d) = \frac{p(c) \sum p(w|c)^{n_i(d)}}{p(d)}$$

AS en twitter

- Se utilizaron 800 tuits, previamente clasificados manualmente.
- De estos, se usaron 80% para entrenamiento y el resto para prueba usando un esquema de Cross Validation.
- El error promedio de entrenamiento fue de $0.192 \pm .015$. El de prueba fue de $0.23 \pm .021$
- Pero...

AS en twitter

- Se utilizaron 800 tuits, previamente clasificados manualmente.
- De estos, se usaron 80% para entrenamiento y el resto para prueba usando un esquema de Cross Validation.
- El error promedio de entrenamiento fue de $0.192 \pm .015$. El de prueba fue de $0.23 \pm .021$
- Pero...

AS en twitter

- Se utilizaron 800 tuits, previamente clasificados manualmente.
- De estos, se usaron 80% para entrenamiento y el resto para prueba usando un esquema de Cross Validation.
- El error promedio de entrenamiento fue de $0.192 \pm .015$. El de prueba fue de $0.23 \pm .021$
- Pero...

AS en twitter

- Se utilizaron 800 tuits, previamente clasificados manualmente.
- De estos, se usaron 80% para entrenamiento y el resto para prueba usando un esquema de Cross Validation.
- El error promedio de entrenamiento fue de $0.192 \pm .015$. El de prueba fue de $0.23 \pm .021$
- Pero...

AS en twitter

Error: 0.195

clase real \ clase estimada	N	O	P
N	14	28	6
O	9	479	63
P	0	15	5

Es necesario buscar opciones para manejar el desbalanceo de categorías.

AS en twitter

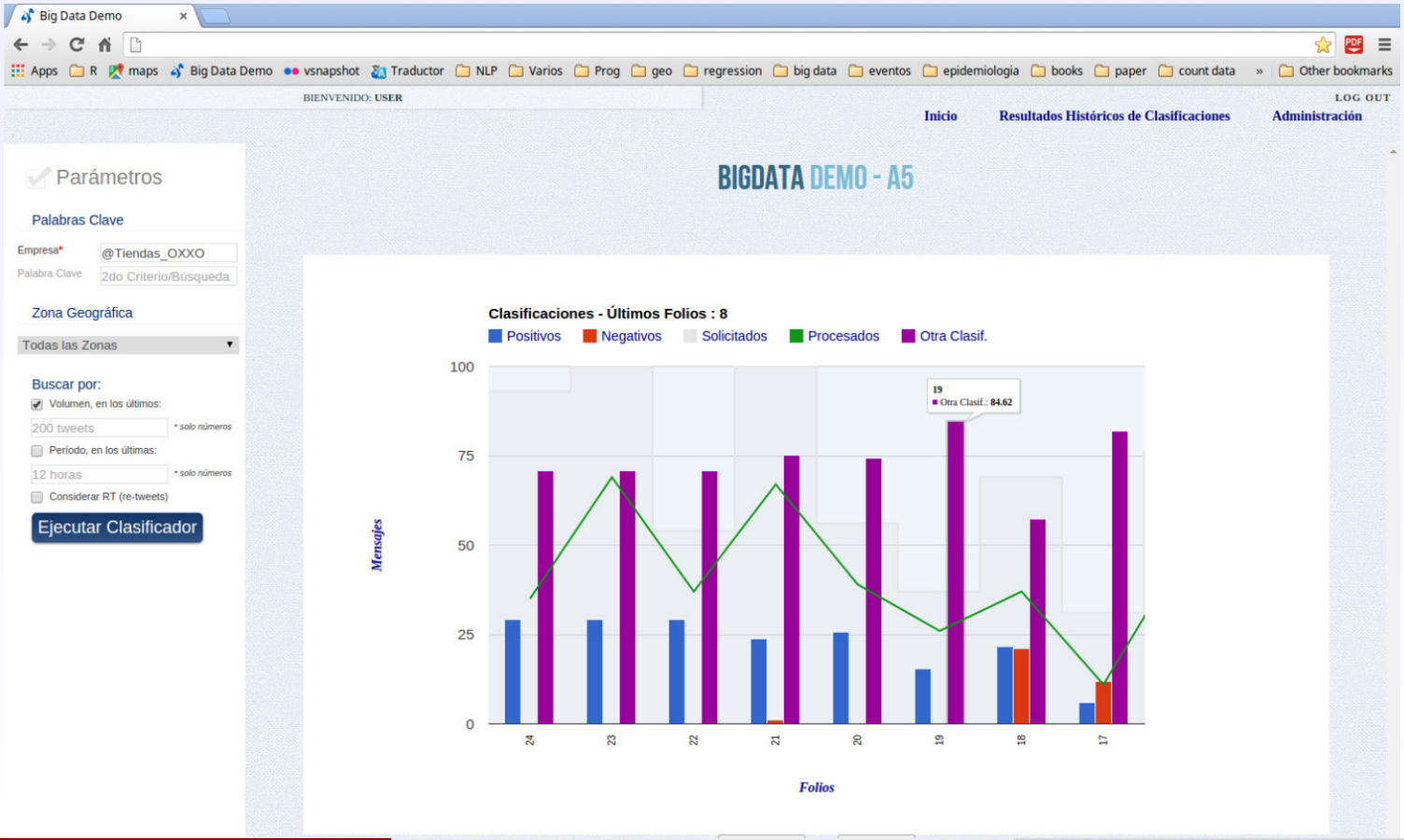
Error: 0.195

clase real \ clase estimada	N	O	P
N	14	28	6
O	9	479	63
P	0	15	5

Es necesario buscar opciones para manejar el desbalanceo de categorías.

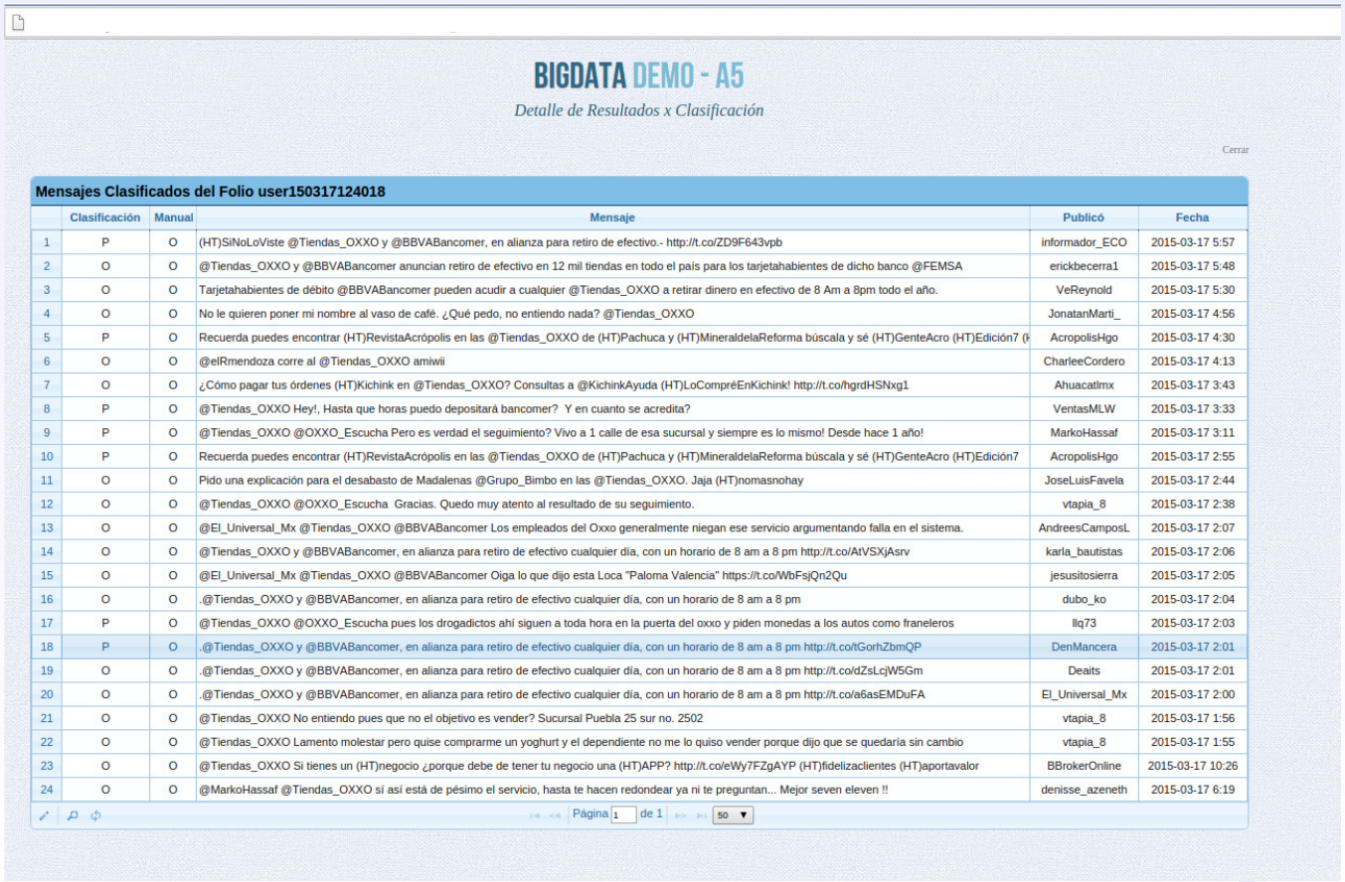
AS en twitter

Desarrollo del sistema de clasificación.
En conjunto con la empresa de consultoría Aleph5, se está desarrollando un sistema de consulta y clasificación de tuits basado en Java y en la API de twitter:



AS en twitter

Desarrollo del sistema de clasificación.
En conjunto con la empresa de consultoría Aleph5, se está desarrollando un sistema de consulta y clasificación de tuits basado en Java y en la API de twitter:



BIGDATA DEMO - A5

Detalle de Resultados x Clasificación

Cerrar

Mensajes Clasificados del Folio user150317124018

	Clasificación	Manual	Mensaje	Publicó	Fecha
1	P	O	(HT)SiNoLoViste @Tiendas_OXXO y @BBVABancomer, en alianza para retiro de efectivo.- http://t.co/ZD9F643vpb	informador_ECO	2015-03-17 5:57
2	O	O	@Tiendas_OXXO y @BBVABancomer anuncian retiro de efectivo en 12 mil tiendas en todo el país para los tarjetahabientes de dicho banco @FEMSA	erickbecerra1	2015-03-17 5:48
3	O	O	Tarjetahabientes de débito @BBVABancomer pueden acudir a cualquier @Tiendas_OXXO a retirar dinero en efectivo de 8 Am a 8pm todo el año.	VeReynold	2015-03-17 5:30
4	O	O	No le quieren poner mi nombre al vaso de café. ¿Qué pedo, no entiendo nada? @Tiendas_OXXO	JonatanMarti_	2015-03-17 4:56
5	P	O	Recuerda puedes encontrar (HT)RevistaAcrópolis en las @Tiendas_OXXO de (HT)Pachuca y (HT)MineraldelaReforma búscala y sé (HT)GenteAcro (HT)Edición7 (H	AcropolisHgo	2015-03-17 4:30
6	O	O	@elRmendoza corre al @Tiendas_OXXO amiwii	CharleeCordero	2015-03-17 4:13
7	O	O	¿Cómo pagar tus órdenes (HT)Kichink en @Tiendas_OXXO? Consultas a @KichinkAyuda (HT)LoCompréEnKichink! http://t.co/hgrdHSNxg1	Ahuacatlmx	2015-03-17 3:43
8	P	O	@Tiendas_OXXO Hey!, Hasta que horas puedo depositará bancomer? Y en cuanto se acredita?	VentasMLW	2015-03-17 3:33
9	P	O	@Tiendas_OXXO @OXXO_Escucha Pero es verdad el seguimiento? Vivo a 1 calle de esa sucursal y siempre es lo mismo! Desde hace 1 año!	MarkoHassaf	2015-03-17 3:11
10	P	O	Recuerda puedes encontrar (HT)RevistaAcrópolis en las @Tiendas_OXXO de (HT)Pachuca y (HT)MineraldelaReforma búscala y sé (HT)GenteAcro (HT)Edición7	AcropolisHgo	2015-03-17 2:55
11	O	O	Pido una explicación para el desabasto de Madalenas @Grupo_Bimbo en las @Tiendas_OXXO. Jaja (HT)nomasnohay	JoseLuisFavela	2015-03-17 2:44
12	O	O	@Tiendas_OXXO @OXXO_Escucha Gracias. Quedo muy atento al resultado de su seguimiento.	vtapia_8	2015-03-17 2:38
13	O	O	@El_Universal_Mx @Tiendas_OXXO @BBVABancomer Los empleados del Oxxo generalmente niegan ese servicio argumentando falla en el sistema.	AndreesCamposL	2015-03-17 2:07
14	O	O	@Tiendas_OXXO y @BBVABancomer, en alianza para retiro de efectivo cualquier día, con un horario de 8 am a 8 pm http://t.co/AtVSXjAsrv	karla_bautistas	2015-03-17 2:06
15	O	O	@El_Universal_Mx @Tiendas_OXXO @BBVABancomer Oiga lo que dijo esta Loca "Paloma Valencia" https://t.co/WbFsjQn2Qu	jesusitosierra	2015-03-17 2:05
16	O	O	.@Tiendas_OXXO y @BBVABancomer, en alianza para retiro de efectivo cualquier día, con un horario de 8 am a 8 pm	dubo_ko	2015-03-17 2:04
17	P	O	@Tiendas_OXXO @OXXO_Escucha pues los drogadictos ahí siguen a toda hora en la puerta del oxxo y piden monedas a los autos como franeleros	llq73	2015-03-17 2:03
18	P	O	.@Tiendas_OXXO y @BBVABancomer, en alianza para retiro de efectivo cualquier día, con un horario de 8 am a 8 pm http://t.co/tGorhZbmQP	DenMancera	2015-03-17 2:01
19	O	O	.@Tiendas_OXXO y @BBVABancomer, en alianza para retiro de efectivo cualquier día, con un horario de 8 am a 8 pm http://t.co/dZsLcjW5Gm	Deaits	2015-03-17 2:01
20	O	O	.@Tiendas_OXXO y @BBVABancomer, en alianza para retiro de efectivo cualquier día, con un horario de 8 am a 8 pm http://t.co/a6asEMDuFA	El_Universal_Mx	2015-03-17 2:00
21	O	O	@Tiendas_OXXO No entiendo pues que no el objetivo es vender? Sucursal Puebla 25 sur no. 2502	vtapia_8	2015-03-17 1:56
22	O	O	@Tiendas_OXXO Lamento molestar pero quise comprarme un yoghurt y el dependiente no me lo quiso vender porque dijo que se quedaría sin cambio	vtapia_8	2015-03-17 1:55
23	O	O	@Tiendas_OXXO Si tienes un (HT)negocio ¿porque debe de tener tu negocio una (HT)APP? http://t.co/eWy7FZgAYP (HT)fidelizaclientes (HT)aportavalor	BBrokerOnline	2015-03-17 10:26
24	O	O	@MarkoHassaf @Tiendas_OXXO sí así está de pésimo el servicio, hasta te hacen redondear ya ni te preguntan... Mejor seven eleven !!	denisse_azeneth	2015-03-17 6:19

Página 1 de 1 50

Conclusiones y trabajo futuro

Conclusiones y trabajo futuro

- El análisis de sentimientos en redes sociales es una tarea complicada
- Aunque hay muchos trabajos publicados, en español hay poco trabajo realizado
- Se requiere mucho preproceso
- Actualmente, se siguen haciendo mejoras en la normalización de los textos

Conclusiones y trabajo futuro

- El análisis de sentimientos en redes sociales es una tarea complicada
- Aunque hay muchos trabajos publicados, en español hay poco trabajo realizado
- Se requiere mucho preproceso
- Actualmente, se siguen haciendo mejoras en la normalización de los textos

Conclusiones y trabajo futuro

- El análisis de sentimientos en redes sociales es una tarea complicada
- Aunque hay muchos trabajos publicados, en español hay poco trabajo realizado
- Se requiere mucho preproceso
- Actualmente, se siguen haciendo mejoras en la normalización de los textos

Conclusiones y trabajo futuro

- El análisis de sentimientos en redes sociales es una tarea complicada
- Aunque hay muchos trabajos publicados, en español hay poco trabajo realizado
- Se requiere mucho preproceso
- Actualmente, se siguen haciendo mejoras en la normalización de los textos

Conclusiones y trabajo futuro

- Se intenta introducir un análisis espacial y temporal de comportamiento en tuits.
- Se explorarán diferentes métodos de clasificación, por ejemplo, basados en SVM, Boosting, y usando String Kernels para los tuits completos.
- Se probarán métodos de clasificación sensibles al costo.

Conclusiones y trabajo futuro

- Se intenta introducir un análisis espacial y temporal de comportamiento en tuits.
- Se explorarán diferentes métodos de clasificación, por ejemplo, basados en SVM, Boosting, y usando String Kernels para los tuits completos.
- Se probarán métodos de clasificación sensibles al costo.

Conclusiones y trabajo futuro

- Se intenta introducir un análisis espacial y temporal de comportamiento en tuits.
- Se explorarán diferentes métodos de clasificación, por ejemplo, basados en SVM, Boosting, y usando String Kernels para los tuits completos.
- Se probarán métodos de clasificación sensibles al costo.

Gracias por su atención!