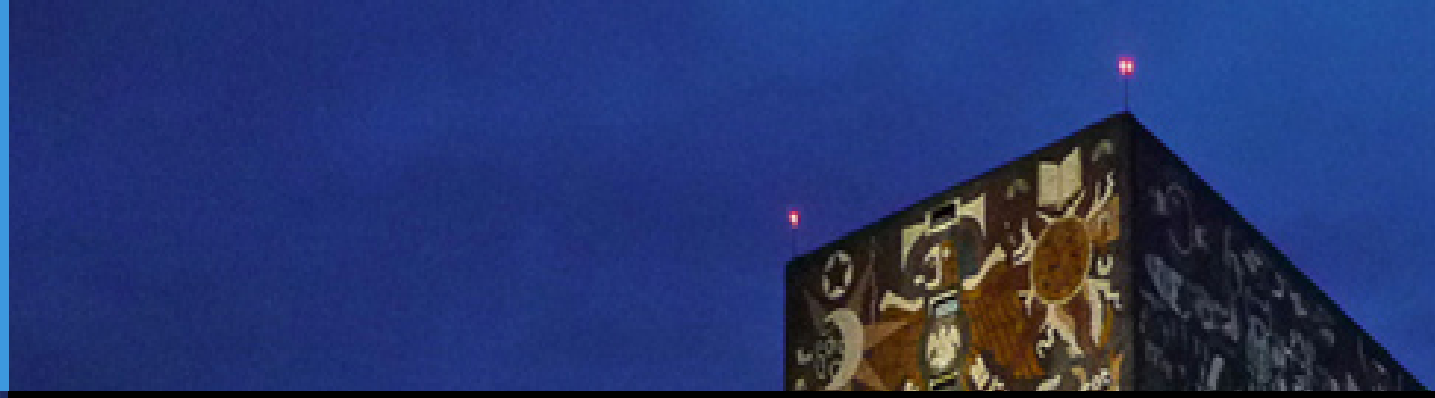
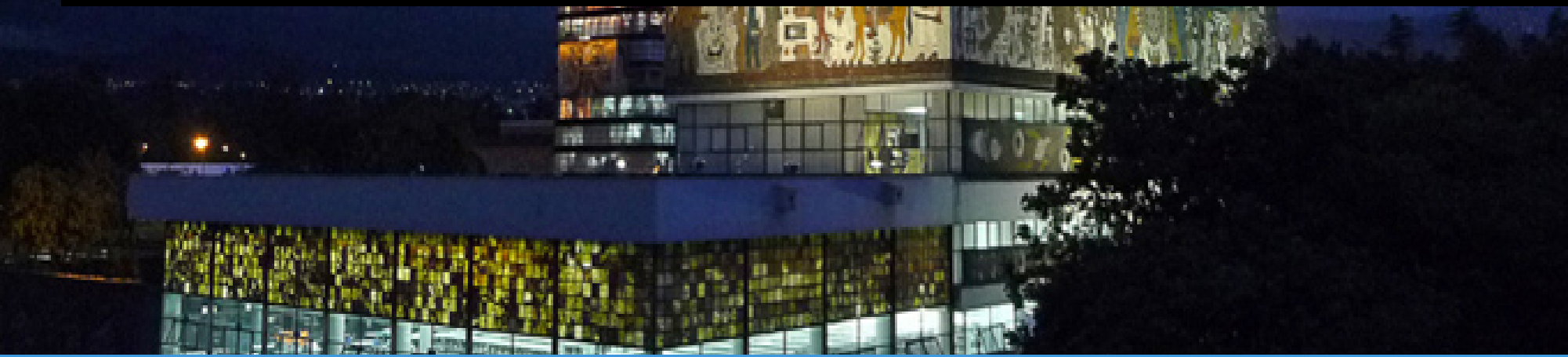


Clasificación automática de la orientación semántica de opiniones mediante bigramas de afirmación y negación

Facultad de Ciencias, UNAM.

Alonso Palomino Garibay

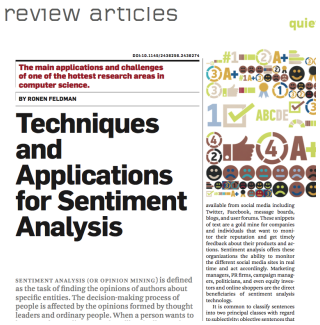
31 de agosto de 2015



Contenidos

Siguiente sección

Introducción - Minería de opiniones


review articles

quie

The main applications and challenges of one of the hottest research areas in computer science.

BY RONEN FELDMAN

Techniques and Applications for Sentiment Analysis

SENTIMENT ANALYSIS (OR OPINION MINING) is defined as the task of finding the opinions of authors about specific entities. The decision-making process of people is affected by the opinions formed by thought leaders and ordinary people. When a person wants to

available from social media including Twitter, Facebook, message boards, blogs, and user forums. Those snippets of text are a gold mine for companies and individuals that want to monitor their reputation and to research consumer feedback about their products and services. Sentiment analysis offers these organizations the ability to monitor the different social media sites in real time and act accordingly. Marketing managers, PR firms, campaign managers, politicians, and even equity investors and online shoppers are the direct beneficiaries of sentiment analysis technology.

It is common to classify sentences into two principal classes with regard to subjectivity: objective sentences that

Figura: Vol. 56 No. 4, Paginas 82-89


A deep, fine-grain analysis of rhetorical structure highlights crucial sentiment-carrying text segments.

BY ALEXANDER HOGENBOOM, FLAVIUS FRASINCAR, FRANCISKA DE JONG, AND UZAY KAYMAK

Using Rhetorical Structure in Sentiment Analysis

POPULAR WEB SITES like Twitter, Blogger, and Epinions let users vent their opinions on just about anything through an ever-increasing amount of short messages, blog posts, and reviews. Automated sentiment-analysis techniques can extract traces of people's sentiment, or attitude toward certain topics, from such texts.[6]

Figura: Vol. 58 No. 7, Paginas 69-77

Figura: *Communications of the ACM*

Introducción - Minería de opiniones

Definición

Minería de opiniones:
Se refiere al estudio computacional de opiniones, sentimientos, evaluaciones, actitudes, apreciaciones, afecciones, puntos de vista, emociones y subjetividades expresadas en texto.

Introducción - Minería de opiniones

LG Electronics 42LF5600 42-Inch 1080p 60Hz LED TV

Add to cart to see price. Why? | In Stock. Ships from and sold by Amazon.com.

Customer Reviews

(26)

4.2 out of 5 stars

5 star		16
4 star		5
3 star		2
2 star		1
1 star		2

Share your thoughts with other customers

Write a customer review

See all 26 customer reviews ›

Most Helpful Customer Reviews

21 of 24 people found the following review helpful

very energy efficient !

By morning fog TOP 1000 REVIEWER VINE VOICE on April 3, 2015

Style Name: TV | Size: 42-Inch

I bought this TV (2015 model) at Target recently, and it has been working well. It only weighs about 20 pounds, so it was easy to carry and to assemble (two legs). The TV was all set after connecting some media boxes and cable and took only about 10 minute. This is a 2015 model, and the design has not changed much from last year's model, but I really like the design. The thin and metallic look bezel (still plastic) is trendy and simple; all thick black bezel looks like a thing of the past. The 2.2 inches of thiness is about half of the thickness from my 55" sammy LED TV, and it makes easy and elegant to hang on the wall. FYI, if one wanted to wall mount, this TV has vesa 400 x 400 mm and comes with a pair of wall mount space for the upper holes.

Introducción - Minería de opiniones

Subtareas dentro de la minería de opiniones:

Introducción - Minería de opiniones

Subtareas dentro de la minería de opiniones:

- Turney (2002)
  - Determinó la orientación semántica a partir de bigramas (¿Positivo o Negativo?).

Introducción - Minería de opiniones

Subtareas dentro de la minería de opiniones:

- Turney (2002)
  - Determinó la orientación semántica a partir de bigramas (¿Positivo o Negativo?).
- Bo Pang et al (2008):
  - Identificación de opiniones, polaridad del sentimiento, resumir de forma automática la orientación de una opinión.

Introducción - Minería de opiniones

Subtareas dentro de la minería de opiniones:

- Turney (2002)
  - Determinó la orientación semántica a partir de bigramas (¿Positivo o Negativo?).
- Bo Pang et al (2008):
  - Identificación de opiniones, polaridad del sentimiento, resumir de forma automática la orientación de una opinión.
- Liu Bing et al (2010)
  - análisis de sentimiento en oraciones de comparación, detección de SPAM, detección de opiniones neutrales y engañosas.

Siguiente sección

Corpus de opiniones

Corpus de trabajo extraído de **ciao.es**[a]

[a] *Sofía N. Galicia-Haro y Alexander Gelbukh (2014).*

Corpus de opiniones

Corpus de trabajo extraído de **ciao.es**[a]

- 2800 opiniones de lavadoras en Español.

[a] *Sofía N. Galicia-Haro y Alexander Gelbukh (2014).*

Corpus de opiniones

Corpus de trabajo extraído de **ciao.es**[a]

- 2800 opiniones de lavadoras en Español.
- Tamaño promedio por lexemas es de 345.

[a] *Sofía N. Galicia-Haro y Alexander Gelbukh (2014).*

Corpus de opiniones

Corpus de trabajo extraído de **ciao.es**[a]

- 2800 opiniones de lavadoras en Español.
- Tamaño promedio por lexemas es de 345.
- El numero total de lexemas de la colección es de 845,280.

[a] *Sofía N. Galicia-Haro y Alexander Gelbukh (2014).*

Corpus de opiniones

Figura: Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0 *(2012)*

Corpus de opiniones

Figura: Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0 *(2012)*

- La colección fue anotada con su lema y categoría gramatical.

Corpus de opiniones

Figura: Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0 *(2012)*

- La colección fue anotada con su lema y categoría gramatical.
- Se utilizaron un conjunto de etiquetas para representar la información morfológica de las palabras.

Corpus de opiniones

Figura: Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0 *(2012)*

- La colección fue anotada con su lema y categoría gramatical.
- Se utilizaron un conjunto de etiquetas para representar la información morfológica de las palabras.
- Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas.

Corpus de opiniones

Corpus de opiniones

- A partir de la colección total de opiniones en Español extrajimos un subconjunto significativo de instancias de opiniones diferentes: 2598.

Corpus de opiniones

- A partir de la colección total de opiniones en Español extrajimos un subconjunto significativo de instancias de opiniones diferentes: 2598.
- Opiniones pagadas por fabricantes

Corpus de opiniones

- A partir de la colección total de opiniones en Español extrajimos un subconjunto significativo de instancias de opiniones diferentes: 2598.
- Opiniones pagadas por fabricantes

Observación

No se eliminaron las opiniones que claramente son anuncios de empresas de mantenimiento (SPAM).

Corpus de opiniones

La tarea para este corpus es la de predicción:

Corpus de opiniones

La tarea para este corpus es la de predicción:

- Determinar qué tan bueno es un producto en base a la orientación semántica de las opiniones de entrenamiento, así como el puntaje de los usuarios.

Corpus de opiniones

La tarea para este corpus es la de predicción:

- Determinar qué tan bueno es un producto en base a la orientación semántica de las opiniones de entrenamiento, así como el puntaje de los usuarios.
- El puntaje de los usuarios que corresponden a: **malo (una estrella)**, **regular (dos estrellas)**, **bueno (tres estrellas)**, **muy bueno (cuatro estrellas)** o **excelente (5 estrellas)**.

Corpus de opiniones

La tarea para este corpus es la de predicción:

- Determinar qué tan bueno es un producto en base a la orientación semántica de las opiniones de entrenamiento, así como el puntaje de los usuarios.
- El puntaje de los usuarios que corresponden a: **malo (una estrella)**, **regular (dos estrellas)**, **bueno (tres estrellas)**, **muy bueno (cuatro estrellas)** o **excelente (5 estrellas)**.
- Errores gramaticales como ortográficos y de puntuación

Corpus de opiniones

La tarea para este corpus es la de predicción:

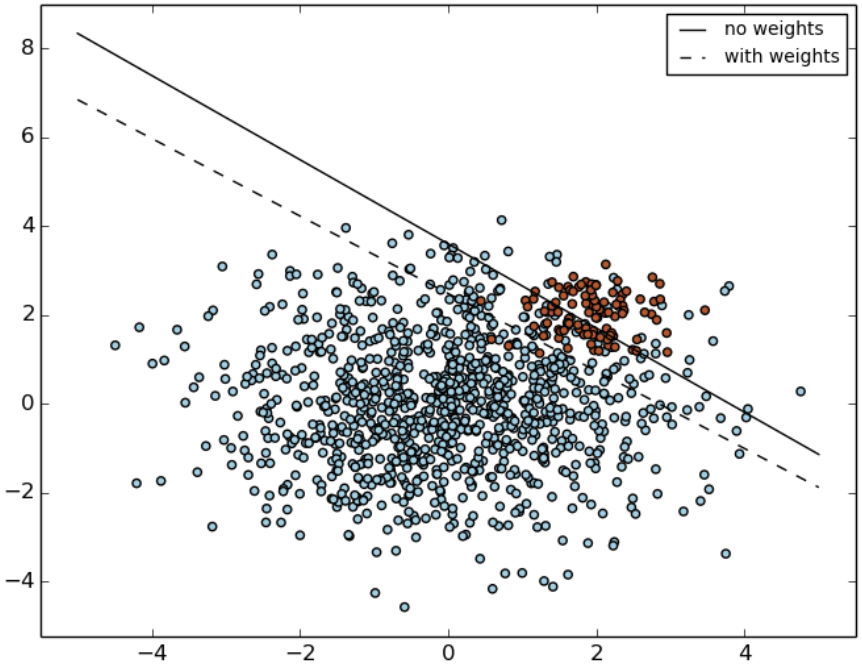
- Determinar qué tan bueno es un producto en base a la orientación semántica de las opiniones de entrenamiento, así como el puntaje de los usuarios.
- El puntaje de los usuarios que corresponden a: **malo (una estrella)**, **regular (dos estrellas)**, **bueno (tres estrellas)**, **muy bueno (cuatro estrellas)** o **excelente (5 estrellas)**.
- Errores gramaticales como ortográficos y de puntuación
- Decidimos no aplicar métodos de corrección automática para normalizar el texto.

Corpus de opiniones

Opiniones	**Detalles**	
Clase	*Numero de instancias*	*Estrellas*
Excelente	1190	5
Muy bueno	838	4
Bueno	239	3
Regular	127	2
Malo	204	1

Figura: Descripción del corpus de reseñas comerciales

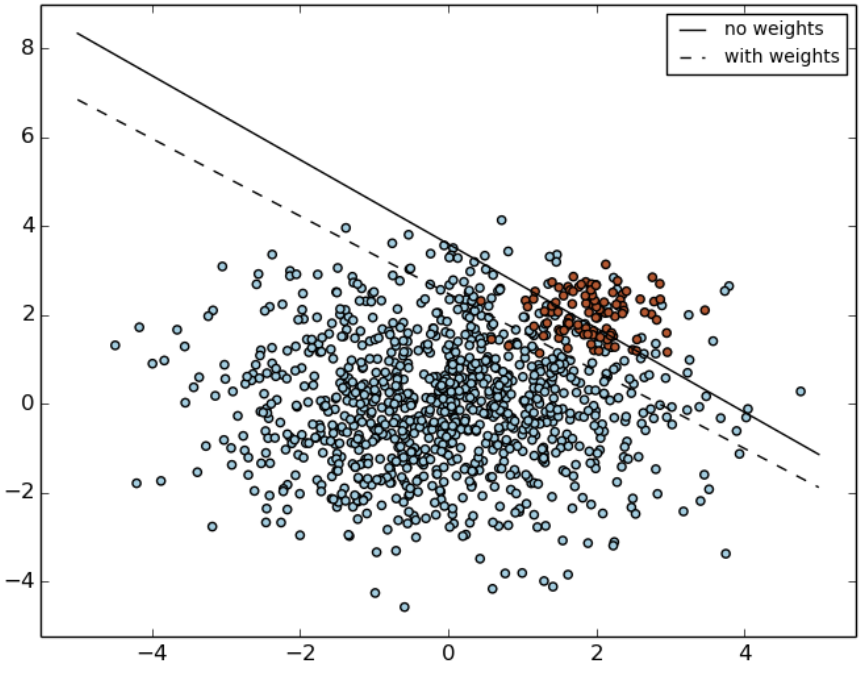
Corpus de opiniones



En el área de aprendizaje automático se ha considerado el problema del desequilibrio de clases.

- Modificación del algoritmo *Sun, Yanmin et al (2007)*.

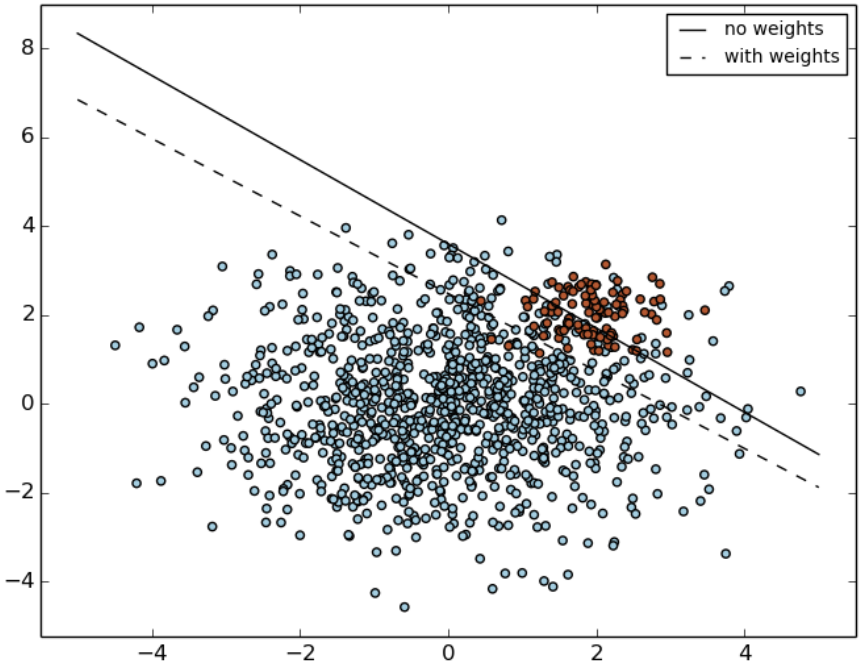
Corpus de opiniones



En el área de aprendizaje automático se ha considerado el problema del desequilibrio de clases.

- Modificación del algoritmo *Sun, Yanmin et al (2007)*.
- Asignación de pesos distintos a los ejemplos de entrenamiento, introduciendo diferentes costos a ejemplos positivos y negativos. *Pazzani, Michael et al (1994)*

Corpus de opiniones



En el área de aprendizaje automático se ha considerado el problema del desequilibrio de clases.

- Modificación del algoritmo *Sun, Yanmin et al (2007)*.
- Asignación de pesos distintos a los ejemplos de entrenamiento, introduciendo diferentes costos a ejemplos positivos y negativos. *Pazzani, Michael et al (1994)*
- Muestreo heterogéneo de datos (e.g. bajo-muestreo, sobre-muestreo, metodos hibridos) *Tang, Yuchun et al (2009)*.

Bigramas

Bigramas

- *Turney, Peter D. (2002)* determinó la orientación semántica mediante una estrategia que consiste en:
  1. Extracción de bigramas a partir de texto.

Bigramas

- *Turney, Peter D. (2002)* determinó la orientación semántica mediante una estrategia que consiste en:
  1. Extracción de bigramas a partir de texto.
  2. Se toman cada bigrama para realizar una búsqueda en la Web empleando el operador NEAR de AltaVista para encontrar cuántos documentos tienen ese bigrama cerca de un término positivo (excellent) y de un término negativo (poor).

Bigramas

- *Turney, Peter D. (2002)* determinó la orientación semántica mediante una estrategia que consiste en:
  1. Extracción de bigramas a partir de texto.
  2. Se toman cada bigrama para realizar una búsqueda en la Web empleando el operador NEAR de AltaVista para encontrar cuántos documentos tienen ese bigrama cerca de un término positivo (excellent) y de un término negativo (poor).
  3. El puntaje para los dos conjuntos se realiza mediante la medida de información mutua puntual *(PMI)*.

Bigramas

- *Turney, Peter D. (2002)* determinó la orientación semántica mediante una estrategia que consiste en:
  1. Extracción de bigramas a partir de texto.
  2. Se toman cada bigrama para realizar una búsqueda en la Web empleando el operador NEAR de AltaVista para encontrar cuántos documentos tienen ese bigrama cerca de un término positivo (excellent) y de un término negativo (poor).
  3. El puntaje para los dos conjuntos se realiza mediante la medida de información mutua puntual *(PMI)*.
     - La diferencia de PMI se utiliza para determinar la orientación semántica

Bigramas

Observación

El puntaje *PMI* de dos palabras $w1$ y $w2$ se obtiene mediante la probabilidad de que las dos palabras aparezcan juntas dividida por la probabilidad de que las dos palabras aparezcan juntas dividida por las probabilidades de cada palabra en forma individual:

$$PMI(w1, w2) = log\left[\frac{P(w1, w2)}{P(w2)P(w2)}\right] \quad (1)$$

Bigramas

La orientación semántica se calculó de la siguiente forma:

Observación

$$SO(frase) = log\left[\frac{hits(Frase\ \mathrm{NEAR}\ excellent)hits(poor)}{hits(frase\ \mathrm{NEAR}\ poor)hits(excellent)}\right] \quad (2)$$

Bigramas

La orientación semántica de bigramas fue utilizada para determinar la orientación semántica de opiniones completas.

Bigramas

La orientación semántica de bigramas fue utilizada para determinar la orientación semántica de opiniones completas.

- Turney tomó 410 comentarios de *epinions.com*

Bigramas

La orientación semántica de bigramas fue utilizada para determinar la orientación semántica de opiniones completas.

- Turney tomó 410 comentarios de *epinions.com*
- Los resultados oscilaron entre el 66 % y 84 % de precisión.

Bigramas

Conclusión

Los bigramas morfosintácticos son una buena característica para métodos no supervisados

Bigramas

Conclusión

Los bigramas morfosintácticos son una buena característica para métodos no supervisados

- Suponemos que para **métodos supervisados** podrían ser mejores.

Bigramas

En este trabajo consideramos los siguientes bigramas morfosintácticos como característica para el entrenamiento del **método supervisado**:

Observación

Estos bigramas morfosintáticos no corresponden a compuestos obtenidos por un analizador sintáctico.

Bigramas

En este trabajo consideramos los siguientes bigramas morfosintácticos como característica para el entrenamiento del **método supervisado**:

Observación

Estos bigramas morfosintáticos no corresponden a compuestos obtenidos por un analizador sintáctico.

- Sustantivo - adjetivo

Bigramas

En este trabajo consideramos los siguientes bigramas morfosintácticos como característica para el entrenamiento del **método supervisado**:

Observación

Estos bigramas morfosintáticos no corresponden a compuestos obtenidos por un analizador sintáctico.

- Sustantivo - adjetivo
- Verbo - adverbio

Bigramas

En este trabajo consideramos los siguientes bigramas morfosintácticos como característica para el entrenamiento del **método supervisado**:

Observación

Estos bigramas morfosintáticos no corresponden a compuestos obtenidos por un analizador sintáctico.

- Sustantivo - adjetivo
- Verbo - adverbio
- Adverbio - adjetivo

Bigramas

En este trabajo consideramos los siguientes bigramas morfosintácticos como característica para el entrenamiento del **método supervisado**:

Observación

Estos bigramas morfosintáticos no corresponden a compuestos obtenidos por un analizador sintáctico.

- Sustantivo - adjetivo
- Verbo - adverbio
- Adverbio - adjetivo
- Adjetivo - adverbio

Bigramas

Mediante un conjunto de scripts a partir de la colección de opiniones, se obtienen todas las secuencias de dos palabras cuyas categorías gramaticales completen los patrones antes indicados (*i.e. bigramas*).

Bigramas

Mediante un conjunto de scripts a partir de la colección de opiniones, se obtienen todas las secuencias de dos palabras cuyas categorías gramaticales completen los patrones antes indicados (*i.e. bigramas*).

- En el caso sustantivo-adjetivo el programa que extrae estos bigramas comprueba la concordancia en género y número.

Bigramas

Mediante un conjunto de scripts a partir de la colección de opiniones, se obtienen todas las secuencias de dos palabras cuyas categorías gramaticales completen los patrones antes indicados (*i.e. bigramas*).

- En el caso sustantivo-adjetivo el programa que extrae estos bigramas comprueba la concordancia en género y número.
- Para todos los bigramas se extraen no solo las palabras, también los lemas.

Bigramas

Mediante un conjunto de scripts a partir de la colección de opiniones, se obtienen todas las secuencias de dos palabras cuyas categorías gramaticales completen los patrones antes indicados (*i.e. bigramas*).

- En el caso sustantivo-adjetivo el programa que extrae estos bigramas comprueba la concordancia en género y número.
- Para todos los bigramas se extraen no solo las palabras, también los lemas.
  - Esto permite agrupar diversas formas en una sola característica.

Bigramas

Mediante un conjunto de scripts a partir de la colección de opiniones, se obtienen todas las secuencias de dos palabras cuyas categorías gramaticales completen los patrones antes indicados (*i.e. bigramas*).

- En el caso sustantivo-adjetivo el programa que extrae estos bigramas comprueba la concordancia en género y número.
- Para todos los bigramas se extraen no solo las palabras, también los lemas.
  - Esto permite agrupar diversas formas en una sola característica.

Ejemplo

Por ejemplo:*prenda vaquera* y *prendas vaqueras*, *lavadora nueva* y *lavadoras nuevas*, se agrupan en un solo bigrama para cada par.

Bigramas

Bigramas

- Bigramas adverbio-adjetivo y adjetivo-adverbio.

Bigramas

- Bigramas adverbio-adjetivo y adjetivo-adverbio.
  - Aunque en Español la forma adverbio-adjetivo es común también encontramos adjetivo-adverbio

Bigramas

- Bigramas adverbio-adjetivo y adjetivo-adverbio.
  - Aunque en Español la forma adverbio-adjetivo es común también encontramos adjetivo-adverbio

Ejemplo

Adjetivo-adverbio: ***poco lento***
Adverbio-adjetivo: ***más eficiente***

Bigramas de Negación

Bigramas de Negación

- La negación esta presente en todos los lenguajes humanos y es usada para revertir la polaridad de un enunciado afirmativo.

Bigramas de Negación

- La negación esta presente en todos los lenguajes humanos y es usada para revertir la polaridad de un enunciado afirmativo.
- Un enunciado negado generalmente tiene implícitamente un significado positivo pero el determinar la parte positiva de la parte negativa de un enunciado es difícil.

Bigramas de Negación

Siguiendo el criterio de [Galicia-Haro *et al* 2015 "Analysis of Negation Cues for Semantic Orientation Classification of Spanish Reviews"] se manejó la negación a nivel de secuencias morfo-sintatcticas y definimos los siguientes patrones:

1. $\text{ninguno}_{\text{LEMMA\_DET}}$ -noun
2. $\text{nada}_{\text{PRONOUN}}$ -adjective
3. [$\text{jamás}_{\text{ADVERB}}$ | $\text{nunca}_{\text{ADVERB}}$ | $\text{no}_{\text{ADVERB}}$]-verb
4. $\text{no}_{\text{ADVERB}}$-$\text{verb}_{\text{AUX\_PAST PARTICIPLE}}$
5. $\text{no}_{\text{ADVERB}}$-pronoun-verb

Siguiente sección

Clasificación

Clasificación

Modelo

Máquinas de soporte vectorial: modelos de aprendizaje supervisado para analizar patrones, usados para clasificación y análisis de regresión.

- Gran variedad de **funciones kernel**.

Clasificación

Modelo

Máquinas de soporte vectorial: modelos de aprendizaje supervisado para analizar patrones, usados para clasificación y análisis de regresión.

- Gran variedad de **funciones kernel**.
- Generalizar en parecencia de muchas. características, usando funciones de nuestro espacio de hipótesis.

Clasificación

Modelo

Máquinas de soporte vectorial: modelos de aprendizaje supervisado para analizar patrones, usados para clasificación y análisis de regresión.

- Gran variedad de **funciones kernel**.
- Generalizar en parecencia de muchas. características, usando funciones de nuestro espacio de hipótesis.
- Uso de heurísticas como **Grid Search** para la optimización de **hiper parámetros**.

Clasificación

Para una tarea de clasificación es necesario separar los datos entre conjunto de entrenamiento y conjunto de prueba:

Clasificación

Para una tarea de clasificación es necesario separar los datos entre conjunto de entrenamiento y conjunto de prueba:

- En nuestro caso separamos el corpus de opiniones en 70 % para entrenamiento y 30 % para prueba.

Clasificación

Para una tarea de clasificación es necesario separar los datos entre conjunto de entrenamiento y conjunto de prueba:

- En nuestro caso separamos el corpus de opiniones en 70 % para entrenamiento y 30 % para prueba.
- Cada ejemplo o instancia se asocia a una clase, categoría o etiqueta

Clasificación

Para una tarea de clasificación es necesario separar los datos entre conjunto de entrenamiento y conjunto de prueba:

- En nuestro caso separamos el corpus de opiniones en 70 % para entrenamiento y 30 % para prueba.
- Cada ejemplo o instancia se asocia a una clase, categoría o etiqueta
  - 70 % de los datos de entrenamiento fueron etiquetados con la clase correspondiente

Clasificación

Para una tarea de clasificación es necesario separar los datos entre conjunto de entrenamiento y conjunto de prueba:

- En nuestro caso separamos el corpus de opiniones en 70 % para entrenamiento y 30 % para prueba.
- Cada ejemplo o instancia se asocia a una clase, categoría o etiqueta
  - 70 % de los datos de entrenamiento fueron etiquetados con la clase correspondiente
  - Mientras que el 30 % de los datos no se les asignó etiqueta.

Siguiente sección

Preprocesamiento de datos

Una de las ventajas de usar un lenguaje de propósito general como Python es la gran cantidad de bibliotecas robustas para implementar distintos métodos y manipular datos.

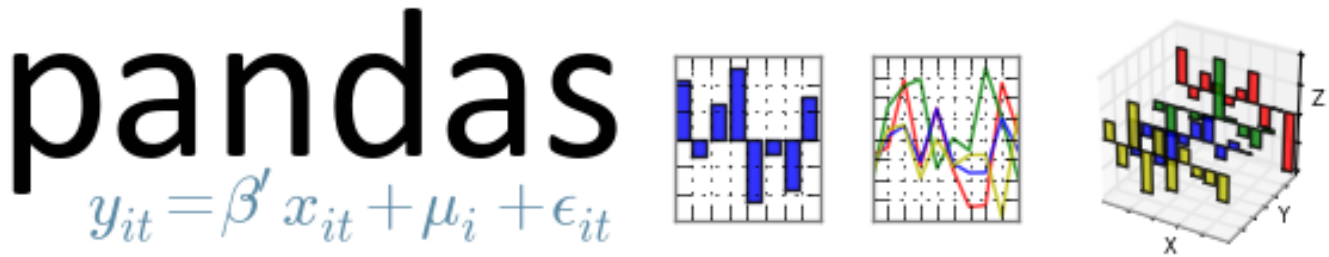


Figura: pandas (Python for data analysis)

Siguiente sección

Sistema

Para resolver este problema de clasificación, decidimos usar un algoritmo supervisado. La clasificación se hizo mediante SVM para el caso multiclase:

Sistema

Para resolver este problema de clasificación, decidimos usar un algoritmo supervisado. La clasificación se hizo mediante SVM para el caso multiclase:

- Fuertes bases teóricas

Sistema

Para resolver este problema de clasificación, decidimos usar un algoritmo supervisado. La clasificación se hizo mediante SVM para el caso multiclase:

- Fuertes bases teóricas
- Algoritmos de aprendizaje que tienen la capacidad de aprender independientemente de la dimensionalidad del espacio de características.

Sistema

Para resolver este problema de clasificación, decidimos usar un algoritmo supervisado. La clasificación se hizo mediante SVM para el caso multiclase:

- Fuertes bases teóricas
- Algoritmos de aprendizaje que tienen la capacidad de aprender independientemente de la dimensionalidad del espacio de características.

Observación

El objetivo de las SVM es producir un modelo basado en los datos de entrenamiento que prediga las clases o categorías de un conjunto nuevo de instancias, mediante la generación de un hiperplano en un espacio de dimensión infinita.

Sistema

Las SVM funcionan para clasificar texto [1]:

[1] *Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. Springer (1998).*

Sistema

Las SVM funcionan para clasificar texto [1]:

- Cuando se clasifica texto se trabaja con espacios de alta dimensión

[1] *Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. Springer (1998).*

Sistema

Las SVM funcionan para clasificar texto [1]:

- Cuando se clasifica texto se trabaja con espacios de alta dimensión
- Pocas características irrelevantes, representaciones vectoriales dispersas

[1] *Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. Springer (1998).*

Sistema

Las SVM funcionan para clasificar texto [1]:

- Cuando se clasifica texto se trabaja con espacios de alta dimensión
- Pocas características irrelevantes, representaciones vectoriales dispersas
- Mayor parte de los problemas de clasificación de texto son linealmente separables.

[1] *Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. Springer (1998).*

Siguiente sección

Experimentos

El entrenamiento de SVM fue realizado empleando la herramienta scikit-learn:

Experimentos

El entrenamiento de SVM fue realizado empleando la herramienta scikit-learn:

- Una biblioteca de código abierto y propósito general.

Experimentos

El entrenamiento de SVM fue realizado empleando la herramienta scikit-learn:

- Una biblioteca de código abierto y propósito general.
- Implementa una gran variedad de algoritmos de aprendizaje automático.

Experimentos

El entrenamiento de SVM fue realizado empleando la herramienta scikit-learn:

- Una biblioteca de código abierto y propósito general.
- Implementa una gran variedad de algoritmos de aprendizaje automático.
- Al igual que otras bibliotecas incorpora o envuelve a la biblioteca de C++ LibSVM.

Siguiente sección

El truco del kernel

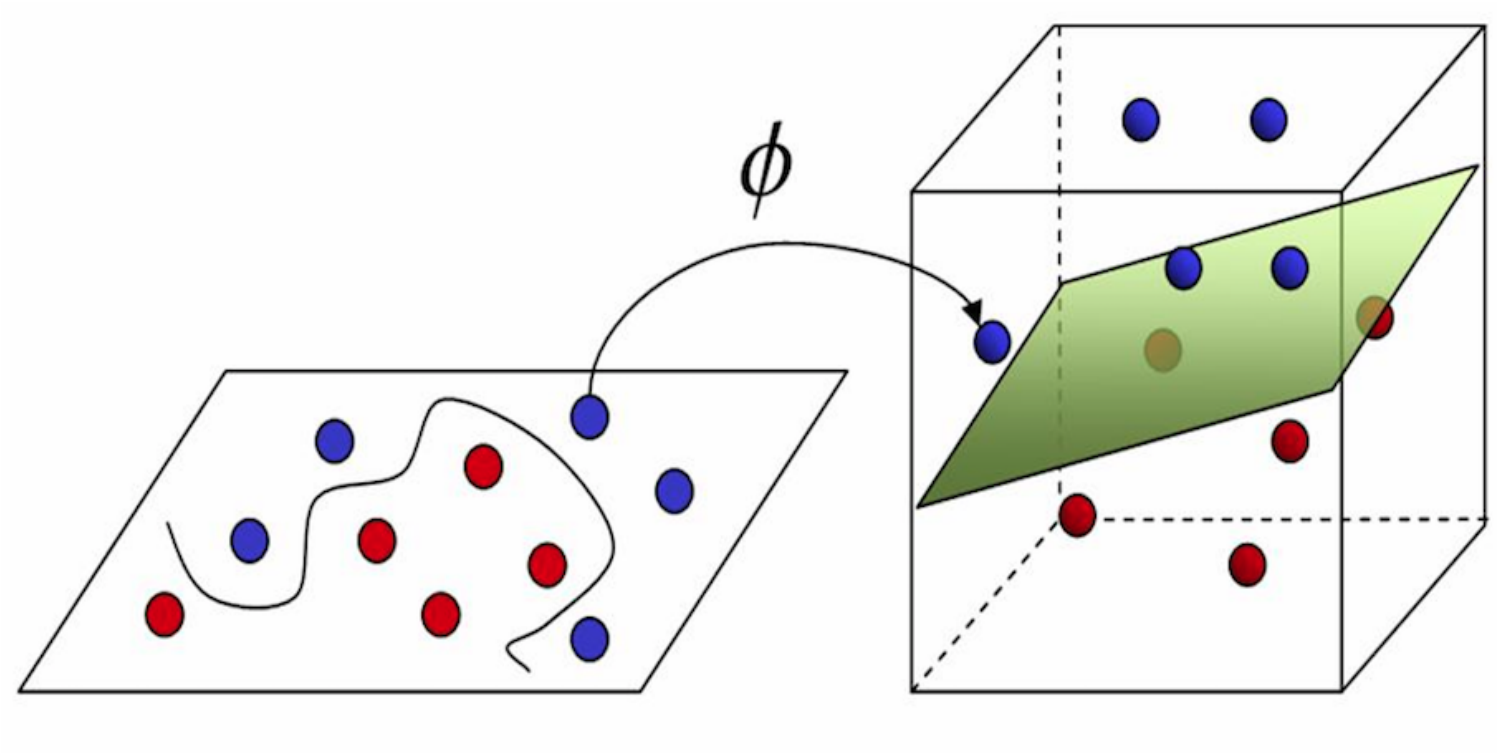


Figura: Truco del kernel

El truco del kernel

Distintas funciones kernel:

El truco del kernel

Distintas funciones kernel:

- RBF (*Función de base radial*)

$$k(x, y) = exp(\gamma||x - y||^2) \tag{3}$$

El truco del kernel

Distintas funciones kernel:

- RBF (*Función de base radial*)

$$k(x, y) = exp(\gamma||x - y||^2) \tag{3}$$

- Kernel polinomial

$$k(x, y) = (\alpha x^\intercal Ty + c)^d \tag{4}$$

El truco del kernel

Distintas funciones kernel:

- RBF (*Función de base radial*)

$$k(x, y) = exp(\gamma||x - y||^2) \tag{3}$$

- Kernel polinomial

$$k(x, y) = (\alpha x^\intercal Ty + c)^d \tag{4}$$

- Kernel lineal

$$k(x, y) = x^\intercal Ty + c \tag{5}$$

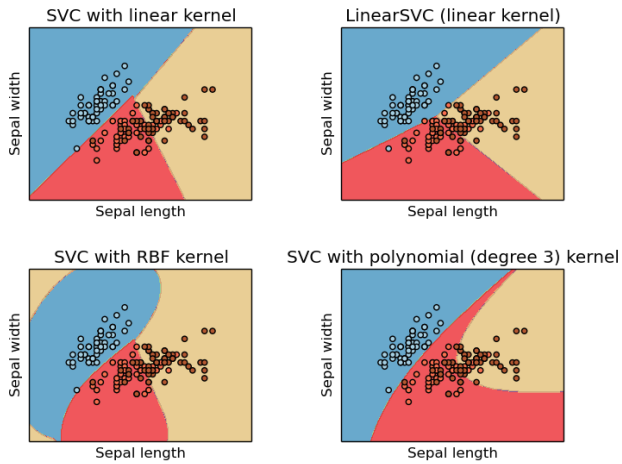
Evaluación

- **Recall:**
  Es la capacidad que tiene un estimador de encontrar todas las muestras positivas. El recall es el radio $\frac{t_p}{t_p+f_n}$ donde t_p es el numero de verdaderos positivos y f_n es el numero de falsos negativos.
- **Precisión:**
  Intuitivamente podemos decir que es la capacidad que tiene un estimador de no etiquetar como positiva una muestra que es negativa. El radio de precisión: $\frac{t_p}{t_p+f_p}$ donde t_p es el numero de verdaderos positivos y f_p el numero de falsos positivos.
- **F1-score:**
  Promedio balanceado entre la precisión y el recall,

Siguiente sección

Grid search

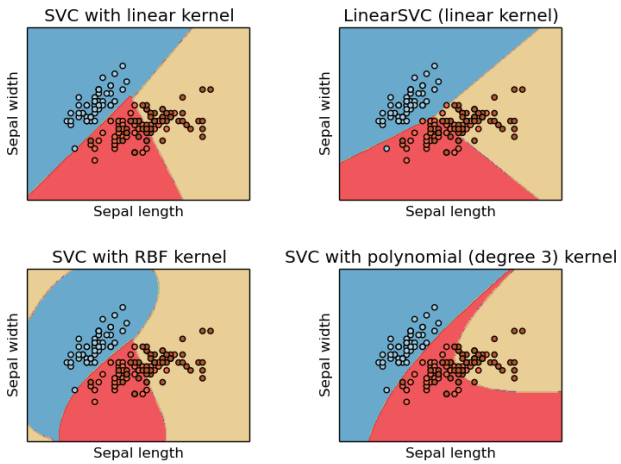
Las SVM son sensibles al conjunto de hiperparametros con las que son entrenados.



Grid search

Las SVM son sensibles al conjunto de hiperparametros con las que son entrenados.

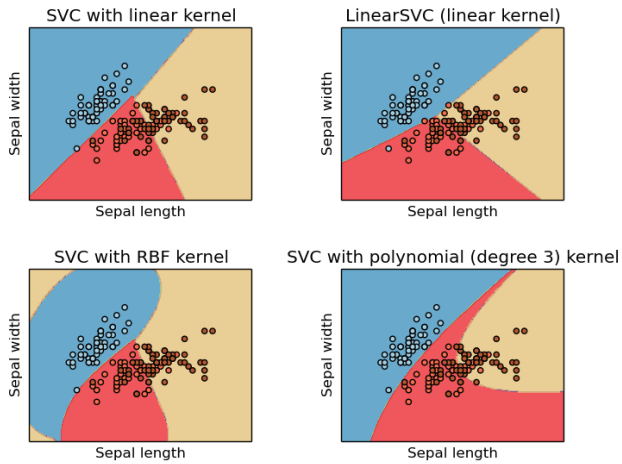
- Un estimador



Grid search

Las SVM son sensibles al conjunto de hiperparametros con las que son entrenados.

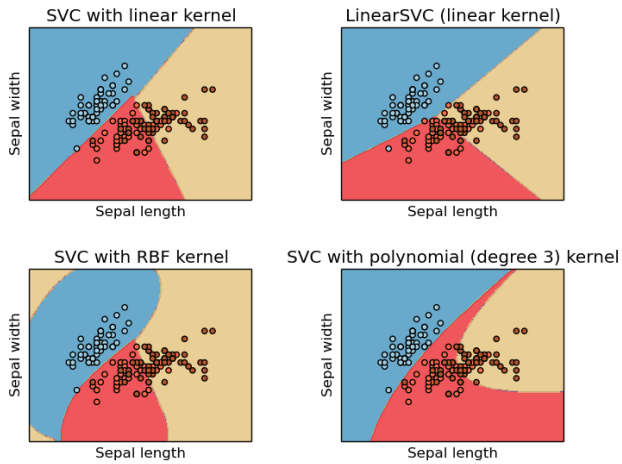
- Un estimador
- Un espacio de parámetros



Grid search

Las SVM son sensibles al conjunto de hiperparametros con las que son entrenados.

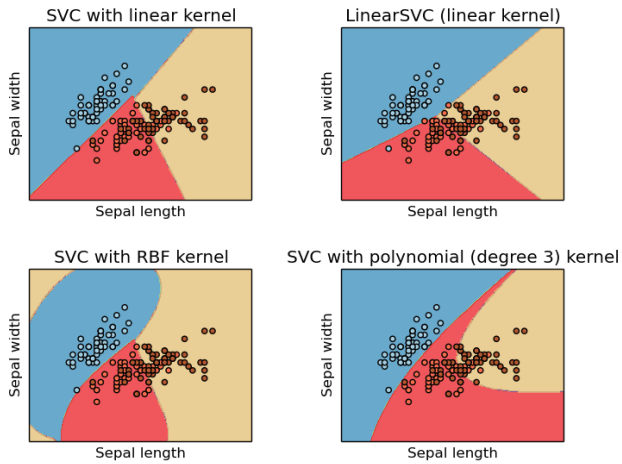
- Un estimador
- Un espacio de parámetros
- Un método para buscar o muestrear candidatos



Grid search

Las SVM son sensibles al conjunto de hiperparametros con las que son entrenados.

- Un estimador
- Un espacio de parámetros
- Un método para buscar o muestrear candidatos
- Un esquema de validación cruzada



Grid search

Observación

Una ***Grid search*** es una búsqueda exhaustiva a través de un subconjunto del espacio de hiper-parámetros de un algoritmo de aprendizaje.

Siguiente sección

Evaluando el rendimiento base

Observación

Evaluar la tasa base de éxito puede aportar un valor mínimo que otro estimador debe superar.(*e.g. tareas de clasificación*).

Evaluando el rendimiento base

Para comparar el resultado usamos un clasificador que usa estrategias simples:

- Es aleatorio.

Evaluando el rendimiento base

Para comparar el resultado usamos un clasificador que usa estrategias simples:

- Es aleatorio.
- Siempre predice la etiqueta más frecuente en el conjunto de entrenamiento.

Evaluando el rendimiento base

Para comparar el resultado usamos un clasificador que usa estrategias simples:

- Es aleatorio.
- Siempre predice la etiqueta más frecuente en el conjunto de entrenamiento.

Observación

Esto es equivalente a usar la estrategia de clasificación más frecuente que implementa la herramienta con la que se hizo el entrenamiento.

Evaluando el rendimiento base

Se obtuvieron los siguientes resultados con el sistema base (*i.e. clasificación más frecuente*):

- *F1 score: 0.33*
- *Recall: 0.33*
- *Precisión: 0.32*

Siguiente sección

Resultados

Generamos distintos conjuntos de entrenamiento:

Resultados

Generamos distintos conjuntos de entrenamiento:

1. Sustantivo-adjetivo
   - Este bigrama expresa atributos sustantivos que corresponden a atributos de características del producto.
   - Exactitud:82.86 y F-beta: 78.22

Resultados

Generamos distintos conjuntos de entrenamiento:

1. Sustantivo-adjetivo
   - Este bigrama expresa atributos sustantivos que corresponden a atributos de características del producto.
   - Exactitud:82.86 y F-beta: 78.22
2. Sustantivo-adjetivo y verbo-adverbio
   - Expresa el modo en que se realiza la acción descrita por el verbo.
   - Mejoró un 10 %
   - Exactitud: 92.65 y F-beta: 92.85

Resultados

Generamos distintos conjuntos de entrenamiento:

1. Sustantivo-adjetivo
   - Este bigrama expresa atributos sustantivos que corresponden a atributos de características del producto.
   - Exactitud:82.86 y F-beta: 78.22
2. Sustantivo-adjetivo y verbo-adverbio
   - Expresa el modo en que se realiza la acción descrita por el verbo.
   - Mejoró un 10 %
   - Exactitud: 92.65 y F-beta: 92.85
3. Sustantivo-adjetivo, verbo-adverbio y adverbio-adjetivo
   - Exactitud: 92.30 y F-beta: 93.23

Resultados

Generamos distintos conjuntos de entrenamiento:

1. Sustantivo-adjetivo
   - Este bigrama expresa atributos sustantivos que corresponden a atributos de características del producto.
   - Exactitud:82.86 y F-beta: 78.22
2. Sustantivo-adjetivo y verbo-adverbio
   - Expresa el modo en que se realiza la acción descrita por el verbo.
   - Mejoró un 10 %
   - Exactitud: 92.65 y F-beta: 92.85
3. Sustantivo-adjetivo, verbo-adverbio y adverbio-adjetivo
   - Exactitud: 92.30 y F-beta: 93.23
4. Sustantivo-adjetivo, verbo-adverbio, adverbio-adjetivo y adjetivo-adverbio
   - No es una estructura lingüística muy usada en Español.
   - *mejor claro*, *super bien*, *perfecto desde_luego*.
   - Exactitud: 93.12 y F-beta: 94.07

Resultados

Table 7. SVM results when negation bigrams were included

Features		Metric	Values
Noun-Adjective	+ no_{ADVERB}-$verb_{AUX\_PAST\ PARTICIPLE}$	F1score	0.9315
		Recall	0.9289
		Precision	0.9342
+	+ $ninguno_{LEMMA\_DET}$ -noun	F1score	0.9406
		Recall	0.9382
		Precision	0.9431
Verb-Adverb	+ $jamás_{ADVERB}$-verb	F1score	0.9380
		Recall	0.9359
		Precision	0.9401
+	+ $nunca_{ADVERB}$ -verb	F1score	0.9524
		Recall	0.9510
		Precision	0.9537
Adverb-Adjective	+ no_{ADVERB} -verb	F1score	0.9407
		Recall	0.9382
		Precision	0.9432
+	+ $nada_{PRONOUN}$ -adjective	F1score	0.9501
		Recall	0.9487
		Precision	0.9515
Adjective-Adverb	+ no_{ADVERB}-pronoun-verb	F1score	0.9395
		Recall	0.9370
		Precision	0.9420

Figura: Rendimiento con bigramas de negación

Resultados

Resultados

- Se han utilizado colecciones de 25 opiniones favorables y 25 opiniones desfavorables para lavadoras con un método no supervisado (*Vilares, David et al 2013*).

Resultados

- Se han utilizado colecciones de 25 opiniones favorables y 25 opiniones desfavorables para lavadoras con un método no supervisado (*Vilares, David et al 2013*).
  - Precisión de 88 para opiniones negativas y 76 para opiniones positivas.

Resultados

- Se han utilizado colecciones de 25 opiniones favorables y 25 opiniones desfavorables para lavadoras con un método no supervisado (*Vilares, David et al 2013*).
  - Precisión de 88 para opiniones negativas y 76 para opiniones positivas.
- Análogamente se han usado SVM para colecciones de opiniones de cine (*Cruz Mata, F. et al 2008*).

Resultados

- Se han utilizado colecciones de 25 opiniones favorables y 25 opiniones desfavorables para lavadoras con un método no supervisado (*Vilares, David et al 2013*).
  - Precisión de 88 para opiniones negativas y 76 para opiniones positivas.
- Análogamente se han usado SVM para colecciones de opiniones de cine (*Cruz Mata, F. et al 2008*).
  - Precisión:87.7, Recall:87.63, F1-Score:87.66

Resultados

- Se han utilizado colecciones de 25 opiniones favorables y 25 opiniones desfavorables para lavadoras con un método no supervisado (*Vilares, David et al 2013*).
  - Precisión de 88 para opiniones negativas y 76 para opiniones positivas.
- Análogamente se han usado SVM para colecciones de opiniones de cine (*Cruz Mata, F. et al 2008*).
  - Precisión:87.7, Recall:87.63, F1-Score:87.66

Conclusión

Estos resultados muestran que el enfoque propuesto en este trabajo se equipara con el estado del arte de minería de opiniones en español.

Siguiente sección

Conclusiones

Conclusiones

- Examinamos el problema de **estimar la orientación semántica de opiniones de productos comerciales**, en **idioma Español**.

Conclusiones

- Examinamos el problema de **estimar la orientación semántica de opiniones de productos comerciales**, en **idioma Español**.
- Exploramos las características de una colección de opiniones

Conclusiones

- Examinamos el problema de **estimar la orientación semántica de opiniones de productos comerciales**, en **idioma Español**.
- Exploramos las características de una colección de opiniones
- Experimentamos con el uso de **bigramas de afirmación** y **bigramas de negación** como **características de entrenamiento** para un **método supervisado** (*Máquinas de soporte vectorial*)

¡Gracias por su atención!