Introducción a la Minería de Información en Redes Sociales

Emmanuel Anguiano Hernández
LTL // CCC // INAOE, Febrero 24 2011

Contenido

- **Minería de texto**

- **¿Qué son las redes sociales?**
  - ¿Porqué se estudia a las redes sociales?

- **¿Cómo se estudia a las redes sociales?**
  - Enfoque Social
  - Minería de Información

- **Twitter**
  - ¿Cómo funciona?
  - Minería de información en twitter

- **Minería de información en twitter**
  - 1. Clasificación de textos cortos en twitter para filtrar información
  - 2. Descubrimiento de conocimiento sobre sentimientos en el flujo de datos de twitter

Minería de datos

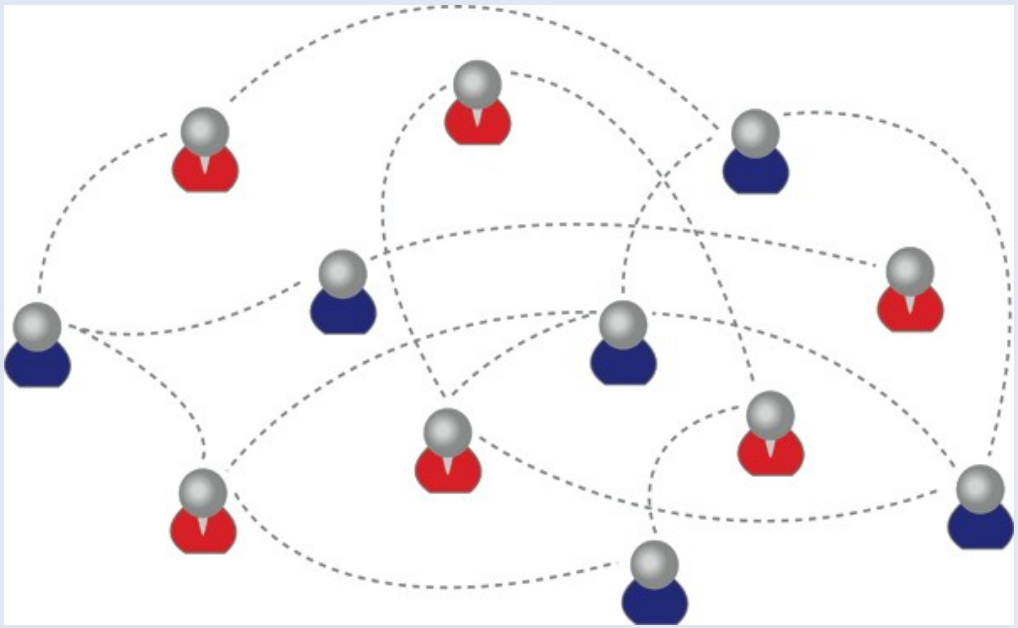
·Es la extracción no trivial de información implícita en un conjunto de datos. Particularmente información de alto nivel a partir de texto.

·Involucra a un conjunto de tareas de NLP:

*Agrupamiento
*Clasificación
*Identificación de entidades
*Extracción de reglas de asociación

¿Qué son las redes sociales?

- **En sociología**: Estructuras sociales formadas por individuos y relaciones que los conectan (amistad, familia, intereses comunes, creencias, conocimiento)



¿Que son las redes sociales?

·**En internet:** Servicios, plataformas o sitios enfocados en reflejar una red o cualquier tipo de relación social entre personas (intereses, preferencias, actividades)

¿Porqué estudiar a las RS?

·La www es una poderosa plataforma para diseminar información, las redes sociales amplifican el efecto.

·Rápido crecimiento

·Gran cantidad de información

¿Cómo se estudia a las RS?

- Perspectiva social
- Minería de información

Tareas desde perspectiva social

- Medición de influencia
- Análisis de comunidades
- Dispersión de la información
- Minería de grafos

Tareas enfocadas en la minería de información

- Minería web
- Clasificación y agrupamiento
- Análisis de opinión
- Identificación de entidades
- Detección temática

Twitter

·”Es una red de información de tiempo real que te permite conectarte con lo que consideres interesante”

·Esta basado en mensajes denominados **tweets**

·La estructura de la red se basa en una relación de **seguimiento**

·La información se proporciona mediante **streams**

Tweet: ¿Qué está pasando?

Es un mensaje de 140 caracteres, puede incluir nombres de @usuario, #hashtags e hipervínculos

fernandeznorona Fernández Noroña
Ayooos ayooooos, me saludan al borracho de @felipecalderon, espero que no haya inventado culpables, especialidad de García Luna
5 hours ago

Tom_Edison Thomas Alva Edison
How Thomas Edison set W. H. Vanderbilt's house on fire - Boing Boing http://j.mp/aphlyL
29 Mar

HomerJSimpson Homer J. Simpson
I wish I had wooden teeth like George Washington. Never have to brush, just paint occasionally.
22 Feb

angelrdzaq Ángel Rodríguez
#pararevivir **#ahorasuena** My generation - The Who ccp @jor_danna
1 hour ago

Twitter: Seguidores

·Un usuario U1 interesado en recibir las publicaciones de otro U2 puede **seguirlo**
·Cada nuevo tweet de U2 aparecerá en la página personal o **timeline** de todos sus seguidores

Twitter: Streams

Un stream es una colección ordenada cronológicamente de los mensajes relacionados con un objeto (usuario, término, hashtag)

Minería de información en twitter

Ventajas:

- Solo texto, longitud uniforme
- Tiempo real, accesible desde cualquier sitio
- Fuerte correlación con noticias de medios *mainstream*
- Potencial para análisis de opiniones

Minería de información en twitter

Dificultades:

- Mucho ruido
- Mensajes muy cortos
- Gran cantidad de información
- Flujo constante

Minería de información en twitter

·**Sentiment Knowledge Discovery in Twitter Streaming Data** (Albert Bifet, Eibe Frank)

·**Sentiment in Twitter Events** (Mike Thelwall, Kevan Buckley, Georgios Paltoglou)

·**Charactering Microblogs with Topic Models** (Daniel Ramage, Susan Dumais, Dan Liebling)

·**From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series** (Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith)

·**Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment** (Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe)

·**Short Text Classification in Twitter to Improve Information Filtering** (Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu, Murat Demirbas)

·**Study of Trend-Stuffing on Twitter through Text Classification** (Danesh Irani, Steve Webb, Calton Pu Kang Li)

·**Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter** (Danah Boyd, Scott Golder, Gilad Lotan)

·**Analysis and Classification of Twitter Messages** (Christopher Horn)

·**NLP-based Approach to Twitter User Classification** (Matt Bush, Ivan Lee, Tony Wu)

Clasificación de Textos Cortos en Twitter para Filtrar Información

Justificación:

Los usuarios fácilmente pueden saturarse de información

Objetivo:

Clasificar automáticamente los tweets recibidos en categorías genéricas: noticias **N**, eventos **E**, opiniones **O**, negocios **D** y mensajes privados **PM**

Método:

Utilizar atributos específicos del dominio

Clasificación de Textos Cortos en Twitter para Filtrar Información

Selección de Atributos:

8 atributos: 1 nominal y 7 atributos binarios

- **Autor** (conocer la fuente)
- **Presencia de abreviaturas o slang** (diferenciar autor corporativo = noticias)
- **Frases con horario de eventos** (eventos)
- **palabras que denotan opinión** (3000 obtenidas de la web)
- **Énfasis en palabras** (caracteres repetidos)
- **Signos de dinero o porcentaje** (negocios)
- **@nom_usuario al principio** (mensajes privados)
- **@nom_usuario en otro sitio** (participación del usuario en eventos)

Clasificación de Textos Cortos en Twitter para Filtrar Información

Coleccion de datos:

-Descargaron al azar, removieron |<3 palabras|, con saludos y |<3 palabras|, con solo una url, con url y |<3 palabras|
-5407 tweets
-684 autores
-Etiquetados manualmente:
  ·2107-N, 625-O, 1100-D, 1075-E, 518-PM
-Removieron stopwords, quedaron 6747 palabras únicas

Clasificador:

Naive Bayes – Validación cruzada de 5 pliegues

Casos Base:

BOW, BOW+autor, 8F, BOW+7F (sin autor), BOW+8F

Clasificación de Textos Cortos en Twitter para Filtrar Información

Resultados:

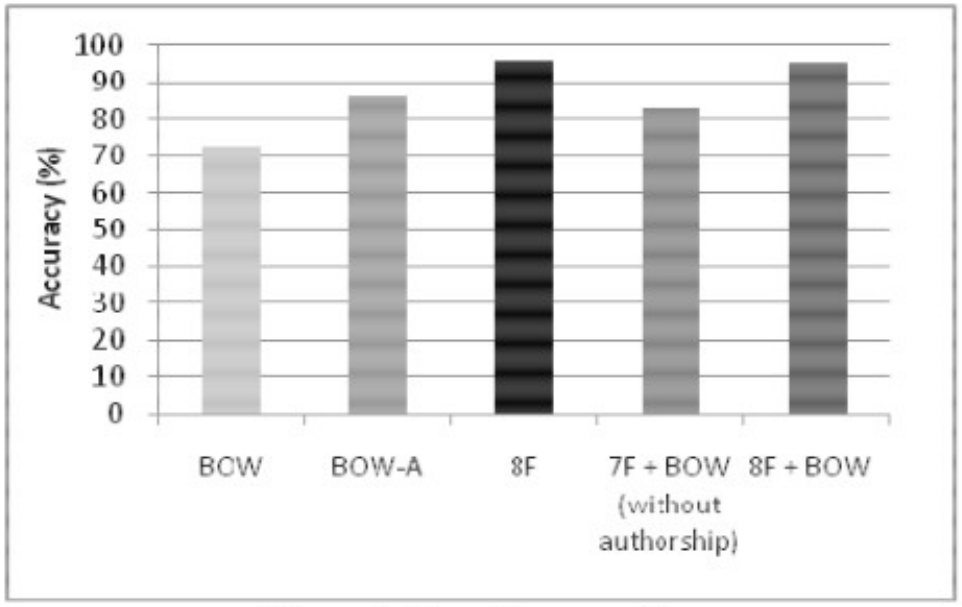


Figure 1. Overall accuracies.

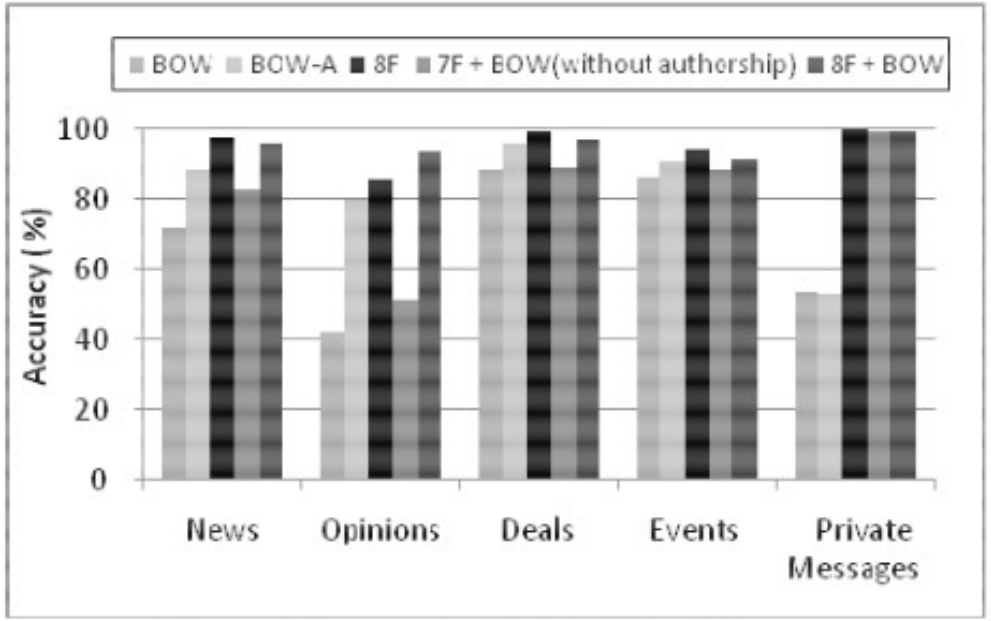


Figure 2. Accuracies for individual classes.

Conclusiones:

- Categorías generales, pero relevantes
- BOW funciona decentemente pero 8F le gana
- Remover ruido podría ser de utilidad

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Introducción:

-Twitter es una potencialmente valiosa fuente de datos que pueden usarse para dar un vistazo en los pensamientos de millones de personas al tiempo que los van manifestando.

-En principio, usar técnicas de minería de flujos podría hacer posible inferir las opiniones de las personas a un nivel tanto individual como grupal sobre cualquier asunto o evento.

-Según el modelo del flujo de datos, los datos llegan a gran velocidad por lo que los algoritmos de minería de datos deben ser capaces de predecir en tiempo real y bajo estrictas restricciones de espacio y tiempo. más aún, deben ser capaces de lidiar con datos cuya naturaleza o distribución cambia en el tiempo.

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

-Evaluar flujos de datos en tiempo real es un reto, hay dos enfoques:

**Conjunto fijo (holdout)*: medir desempeño usando un conjunto fijo

**Intercalado (interleaved) test-then-train o precuencial*: cada muestra individual es usada para probar el modelo antes de usarla para entrenamiento, la exactitud es incrementalmente actualizada.

-Los flujos de datos suelen estar desbalanceados, por lo que un 90-10 podría alcanzar 90% de exactitud arbitrariamente. Se propone el uso de la estadística Kappa basada en una ventana deslizante para medir el desempeño

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Minando Datos de Twitter: Retos y Trabajo Relacionado

-En un contexto de descubrimiento de conocimiento hay dos tareas que pueden considerarse en conjunto con los datos de twitter:

**Minería de grafos (análisis de vínculos)*

- ·Medir la influencia y dinámica de la población
- ·Descubrimiento y formación de comunidades
- ·Difusión social de la información

**Minería de texto (análisis del contenido de los mensajes)*

- ·Análisis de sentimientos
- ·Clasificación, agrupamiento
- ·Detección de tópicos

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Evaluación de Flujos de Datos con Clases Desbalanceadas

-La medida de evaluación más usada es la exactitud, pero solo es válida cuando las clases están balanceadas.

-Se propone la estadística Kappa como una medida más sensible.

-Solo la evaluación precuencial provee una imagen del comportamiento en el tiempo pero no es adecuada para datos desbalanceados por lo que se propone una 'estimación precuencial de Kappa'.

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table 1. Simple confusion matrix example

	Predicted Class+	Predicted Class-	Total
Correct Class+	68.06	14.94	83
Correct Class-	13.94	3.06	17
Total	82	18	100

Table 2. Confusion matrix for chance predictor based on example in Table 1

-La estadística kappa normaliza la exactitud de un clasificador con la de un predictor aleatorio.

-Es particularmente útil en minería de flujos debido a los cambios potenciales en la distribución de las clases. si el clasificador está siempre en lo correcto k=1, si las predicciones coinciden con las del predictor aleatorio entonces k=0.

-Cómo contar los objetos relevantes?: con una ventana deslizante con las w más recientes observaciones.

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Métodos Para Minería De Flujos De Datos

-Se realizan experimentos con tres métodos incrementales rápidos:

-***Naïve Bayes Multinomial*:** el documento es una bolsa de palabras y hay que estimar la probabilidad condicional de que un documento pertenezca a una clase dados los términos que aparecen en él.

-Descenso de Gradiente Estocástico: provee una manera eficiente de aprender algunos clasificadores, aún si están basado en una función de perdida no diferenciable como la 'hinge loss' usada en SVM.

***-Árbol Hoeffding*:** emplea una estrategia de pre-poda basada en el límite Hoeffding para hacer crecer incrementalmente un árbol de decisión. Un nodo es expandido por división tan pronto como hay evidencia estadística basada en los datos observados.

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Evaluación Experimental

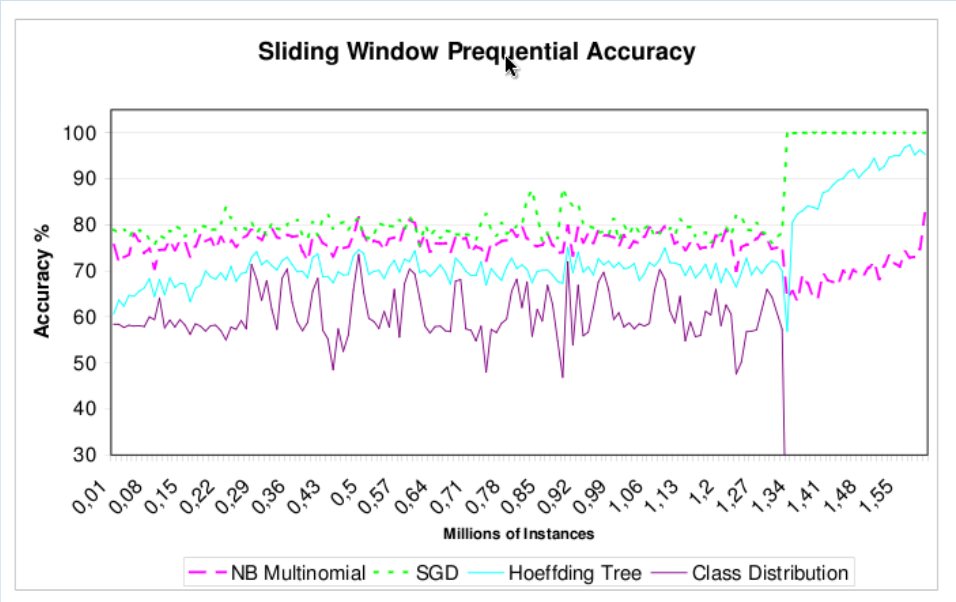
-MOA (Massive Online Análysis) es un sistema para aprendizaje a partir de muestras como los flujos de datos. Los algoritmos fueron implementados en java usando Weka y MOA.
-Los tweets fueron representados como un conjunto de palabras usando 10000 unigramas extraídos del conjunto de entrenamiento. Se usó presencia del término.

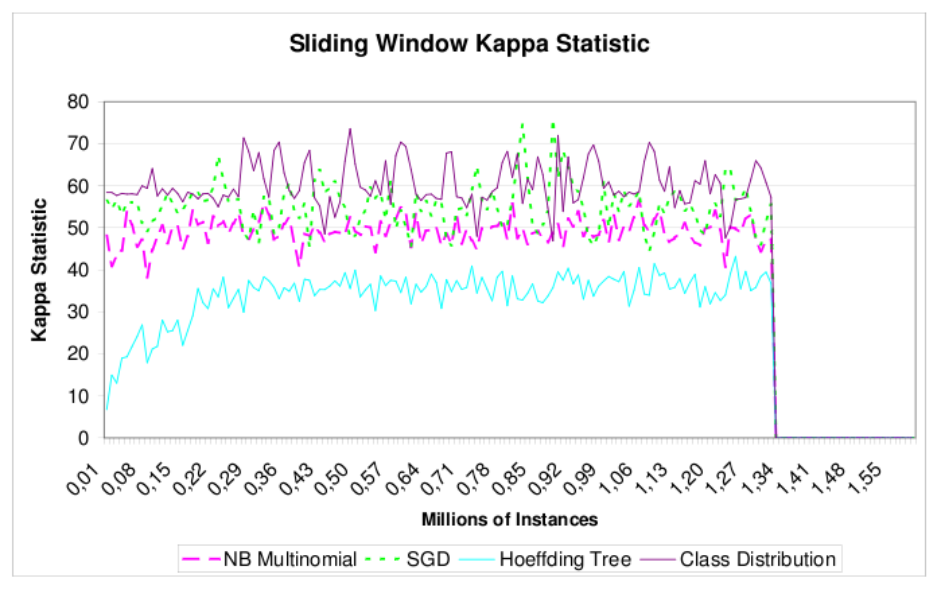
-twittersentiment.appspot.com: permite al usuario rastrear los sentimientos sobre una marca, producto o tópico. en su desarrollo usaron un conjunto de entrenamiento con 800,000 t que contenían emoticones positivos y la misma cantidad para negativos, el conjunto de prueba estuvo formado por 182 t+ y 177 t- etiquetados manualmente.
-Corpus de Edimburgo: es una colección de ts recolectados entre nov11-2009 y feb1-2010, contiene 97 millones de ts multilingües, solo se usaron los tweets en inglés que contenían emoticonos: 324,917 t- y 1,813,705 t+.

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Resultados

- **Experimento 1:** twittersentiment.appspot.com, ventana de 1000t.





	Accuracy	Kappa	Time
Multinomial Naïve Bayes	75.05%	50.10%	116.62 sec.
SGD	82.80%	62.60%	219.54 sec.
Hoeffding Tree	73.11%	46.23%	5525.51 sec.

Table 3. Total prequential accuracy and Kappa measured on the `twittersentiment.appspot.com` data stream

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Resultados

- **Experimento 2:** twittersentiment.appspot.com, modo clásico.

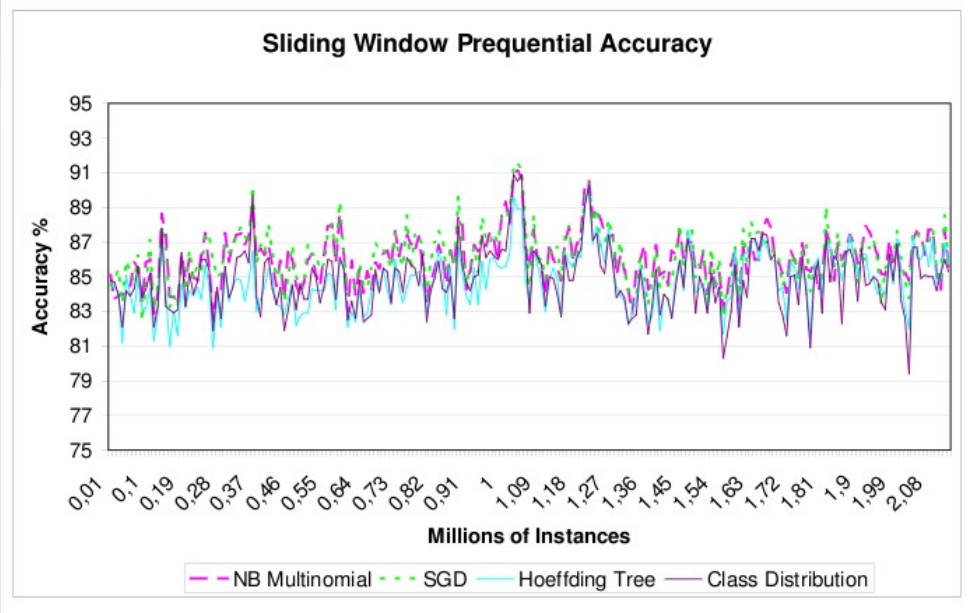
	Accuracy	Kappa
Multinomial Naïve Bayes	82.45%	64.89%
SGD	78.55%	57.23%
Hoeffding Tree	69.36%	38.73%

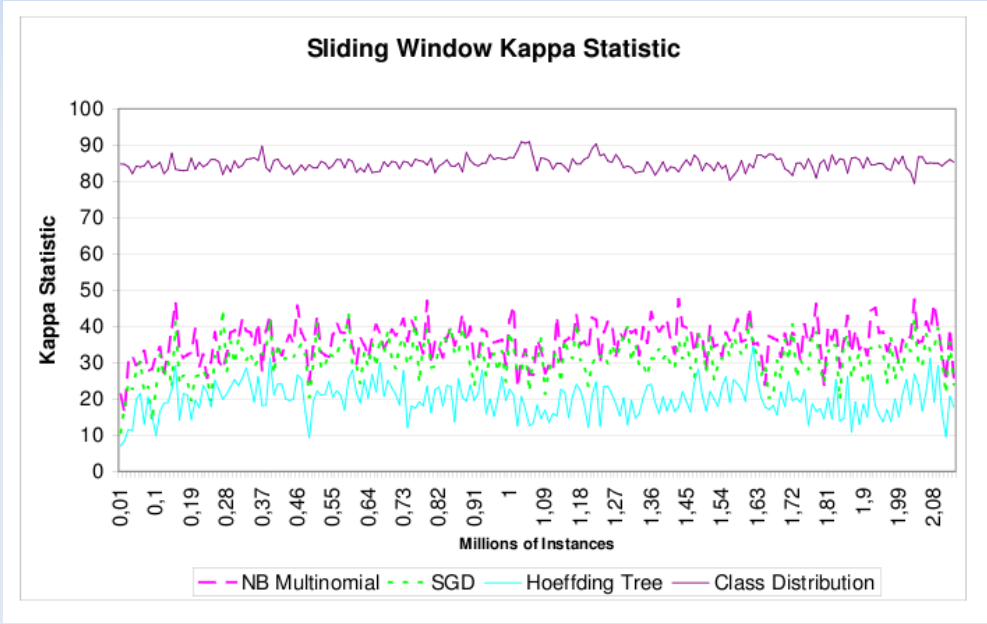
Table 4. Accuracy and Kappa for the test dataset obtained from `twittersentiment.appspot.com`

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Resultados

·**Experimento 3:** Corpus de Edimburgo, ventana de 1000t.





	Accuracy	Kappa	Time
Multinomial Naïve Bayes	86.11%	36.15%	173.28, sec.
SGD	86.26%	31.88%	293.98 sec.
Hoeffding Tree	84.76%	20.40%	6151.51 sec.

Table 5. Total prequential accuracy and Kappa obtained on the Edinburgh corpus data stream.

Descubrimiento de Conocimiento sobre Sentimientos en el Flujo de Datos de Twitter

Conclusiones

-Twitter tiene la capacidad potencial para permitir a cualquier usuario saber qué está pasando en el mundo en cualquier momento de tiempo.

-Las técnicas de minería de flujos son las que mejor encajan con la naturaleza de twitter (Straming API) y no habían sido consideradas.

-Se propuso la estadística Kappa de ventana deslizante como métrica de evaluación.

-De acuerdo con los resultados de las pruebas, el modelos basado en SGD se recomienda para este tipo de datos.

Gracias.