Artículo **Revista de Tecnologías de la Información**
Marzo 2015 Vol.2 No.2 90-103

Explotación de información de redes sociales de microbloggs utilizando análisis de sentimientos y técnicas propuestas por el big data

GONZALEZ-MARRON, D.†, MEJIA-GUZMAN, D., GONZALEZ-MENDOZA, M. y ENCISO-GONZALEZ, A.

Instituto Tecnológico de Pachuca

Recibido 19 de Enero, 2015; Aceptado 10 de Marzo, 2015

Resumen

En este artículo se muestra la implementación de una aplicación, que permite extraer información de la red social de twitter, utilizando como primer objetivo el análisis de polaridad del mensaje, comúnmente denominado como análisis de sentimientos (AS), se detalla una estrategia para seleccionar el algoritmo a utilizar y los resultados obtenidos al seleccionar un tema específico de la red social, se describen también las ventajas que puede proporcionar el uso de técnicas de big data a esta aplicación.

Redes sociales, twitter, big data, análisis de sentimientos

Abstract

In this paper an application that extracts information of the twitter social network is presented, the first obtective is to determine the message polarity of twitter messages using sentiment analysis (SA) techniques. It is explained the strategy to select the best algorithm to realize SA. The results obtained for specific queries of themes selected are reported, also the advantages that big data architecture gives to this application are mentioned.

Social Netwoks, Twitter, Big data, sentiment analysis

Citación: GONZALEZ-MARRON, D., MEJIA-GUZMAN, D., GONZALEZ-MENDOZA, M. y ENCISO-GONZALEZ, A. Explotación de información de redes sociales de microbloggs utilizando análisis de sentimientos y técnicas propuestas por el big data. Revista de Tecnologías de la Información 2015, 2-2: 90-103

† Investigador contribuyendo como primer autor.

Introducción

En la actualidad se está produciendo mucha más información que en años anteriores, debido principalmente a la automatización de la información, al incremento en las unidades de almacenamiento, y a una disminución de costos, es por ello que los usuarios en la actualidad tienden a acumular cada dia mas datos y esta tendencia continua en ascenso.

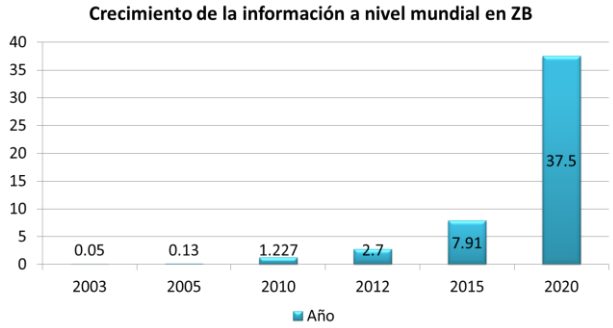


Figura 1 Crecimiento de información en Zbytes

Es importante resaltar que la evolución en capacidades de almacenamiento de los dispositivos electrónicos, se ha incrementado en 50,000,000 de veces desde sus origenes, en sus inicios las capacidades de almacenamiento en una pulgada cuadrada eran de 2000 bits y ahora las capacidades son de 100 gigabits, ésta es sin duda la variable de las computadoras que más se ha mejorado, causando que cada vez se pueda almacenar más información y que este almacenamiento sea más barato, sin embargo este incremento ha hecho que cada día sea más difícil el procesamiento de este gran volumen de datos.

Es por eso que en la actualidad es necesario considerar técnicas que permitan procesar una mayor cantidad de datos, como el big data que procesa grandes volúmenes, haciendo un procesamiento distribuido y en paralelo. En este artículo se detalla el procesamiento de información requerido para procesar los volúmenes de información generada[1].

Cabe mencionar que se seleccionó la red de twitter como la fuente de información para hacer análisis debido a que se conforma de mensajes con un tamaño máximo de 140 caracteres.

Extracción de información de redes Sociales

Una empresa o institución más que ofrecer un producto o servicio, ofrece una experiencia, y muchas de éstas son narradas, compartidas y leídas por los usuarios en las redes sociales.

Una red social es definida como una estructura social conformada por organizaciones o individuos con un interés común, que permite a éstos crear un perfil público o semipúblico en un sistema delimitado, además, permite conformar una lista de otros usuarios con los que se comparte una conexión, y ver y recorrer su lista de conexiones y de las realizadas por otros dentro del sistema[2]. En la actualidad es posible extraer información de utilidad para la mejora en toma de decisiones de las redes sociales, algunos sitios de internet como Facebook, Twitter, Google plus, entre otros ofrecen plataformas para que desarrolladores independientes puedan usar información de dichos sitios y crear nuevas aplicaciones. El poder tener información de redes sociales permite hacer estudios minuciosos del comportamiento de los usuarios ante diferentes sucesos para poder llegar al conocimiento y posteriormente al metaconocimiento, como puede ser visto en la siguiente figura.

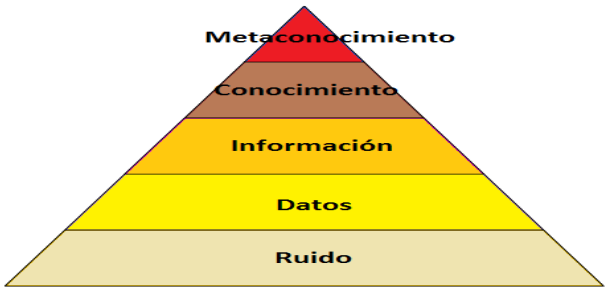


Figura 2 Jerarquía del Conocimiento

ISSN-2410-4000

Según el portal de tráfico Alexa[3], las redes sociales mas populares en el mundo son Facebook, Twitter, Linkedin, Google plus.

De estas redes sociales, fue seleccionada la red de twitter, debido a que la información está organizada en mensajes con una longitud máxima de 140 caracteres, lo cual la hace una fuente de tipo homogéneo, a diferencia de las otras redes que constan de diferentes fuentes de información, como son: video, imágenes y sonidos, lo que dificulta su procesamiento. Twitter refleja mejor las tendencias actuales de tópicos seleccionados por los usuarios, comparado con otras redes.

Big data y las bases de datos NoSQL [8].

Big Data cuyo significado textual se refiere a grandes volúmenes de datos, considera igualmente la velocidad requerida para procesar los datos y la variedad de éstos. La convergencia de estas tres V´s define la caracterización primitiva de Big Data.

El volumen permite caracterizar grandes colecciones de datos creadas para diferentes usos y propósitos. El almacenamiento de Big Data es el reto más inmediato, ya que la primera responsabilidad es la de preservar todos los datos generados en el ámbito de actuación del sistema. La decisión de cómo se almacenan los datos tiene, a su vez, un impacto considerable en el rendimiento de los procesos de recuperación, procesamiento y análisis de Big Data.

La velocidad caracteriza los flujos de datos desarrollados en entornos cada vez más distribuidos.

Se pueden distinguir dos tipos: flujos de nuevos datos (generados de diferentes maneras y por diferentes fuentes) que deben ser integrados de forma progresiva en los Big Data existentes y flujos que contienen los resultados de las consultas y cuyo volumen puede ser potencialmente grande. Por lo tanto, la velocidad describe lo rápido que se generan, demandan y entregan los datos en su entorno de explotación.

La variedad se refiere a los diferentes grados de estructura (o falta de ella) que pueden encontrarse en una colección datos que puede integrar procedentes de múltiples fuentes (redes de sensores, logs generados en servidores web, redes sociales, datos de origen político, económico o científico, entre otros) y, obviamente, cada una de ellas posee una semántica particular que da lugar a esquemas diferentes que son difícilmente integrables en un modelo único.

Cualquier arquitectura diseñada para la gestión de Big Data debe integrar las tres dimensiones anteriores.

Para este caso la información que se utiliza proviene de una red social, la cual deberá ser consultada para analizar sus datos, se debe tener acceso a la información, y al mismo tiempo, se deberá contar con capacidad y medios para descifrarla, interpretarla y encontrar las correlaciones que se esconden en su contenido.

Anteriormente, con la información que se tenía solo se podía saber lo que pasó, pero ahora interesa más conocer lo que pasará; y todos estos aspectos requieren de nuevos modelos de análisis mucho más complejos como lo son en la actualidad los utilizados por la inteligencia de negocios.

ISSN-2410-4000

Es por ello que muchas de las decisiones que se toman, derivan en acciones que se obtienen como resultado de la implementación de un proyecto de Big Data.

La implementación desarrollada parte de datos no estructurados (obtenidos de la red social Twitter o Facebook), es por ello que los resultados se sustentan no solo en el volumen, sino también en el poder de analizarlos. La arquitectura en la que se basa este desarrollo es la arquitectura Lambda, cuyo modelo fusiona el procesamiento por lotes y el procesamiento en tiempo real. Se Compone de 3 capas:

- Batch layer
- Serving layer
- Speed layer

La capa por lotes (Batch layer) aborda el almacenamiento de Big Data e implementa los mecanismos para la construcción de diferentes tipos de vistas sobre los datos, (batch views). La capa de Servicio (Serving Layer), carga estas vistas para su consulta, mientras que la capa de Filtrado (Speed Layer) gestiona las necesidades del tiempo real mediante procesos de actualización incremental de los datos que almacena y de las vistas que construye sobre ellos. Esta capa se responsabiliza de recoger los datos generados en tiempo real y almacenarlos temporalmente hasta que se integren en el Big Data residente en la capa Batch. Una vez que se decide esta integración, se generan desde cero todas las vistas de la capa de Servicio y se eliminan todos los datos en la capa de Velocidad.

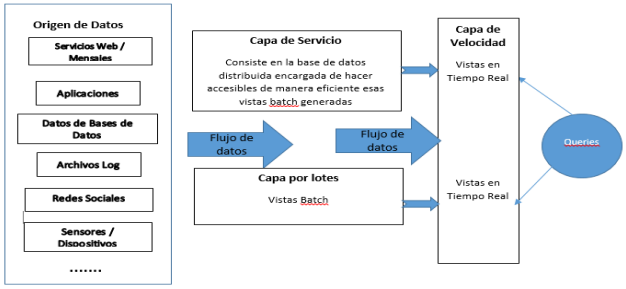


Figura 3 Arquitectura Lambda

Basada en HDFS (Hadoop Distributed File System):

Se compone de 2 elementos:

- NameNode: Servidor que gestiona el espacio de nombres del sistema de archivos y regula el acceso a los archivos de los clientes, por cada clúster HDFS existe uno solo.
- DataNodes: Nodos que gestionan el almacenamiento de datos.

Cabe recalcar que para este proyecto la conceptualización dada por Houghton Mifflin Harcourt, 2013 *"Big Data: Una revolución que transformará cómo vivimos, trabajamos y pensamos"*, refleja la temática de este proyecto.

Principales diferencias entre las bases de datos NoSQL con las relacionales

- No utilizan lenguaje SQL
- No permiten el uso de JOIN
- No garantizan la propiedad ACID
- Escalabilidad horizontal

Teorema CAP para las Bases de datos NoSQL

Las bases de datos NoSQL, tienen la característica de ser escalables horizontalmente y para ello, se deben implementar en sistemas distribuidos, el Teorema CAP fue desarrollado por Eric Brewer, quien declara que un sistema distribuido puede cumplir solo con 2 de las tres siguientes características:

- Consistencia (Consistency): Aseguramiento de datos completos y correctos.

- Disponibilidad (Availability): La información de la base de datos siempre esta disponible.

- Tolerancia de las Particiones (Partition Tolerance): Las bases de datos deben funcionar de manera correcta aunque existan problemas de conexión.

- Por lo tanto existen 3 combinaciones en las que puede clasificar una base de datos NoSQL.

    - C y A: Sistemas en los que si falla la comunicación entre sus nodos el conjunto no puede trabajar.

    - C y P: En este tipo de sistemas, si ocurre un incidente, parte de la información no estará disponible, pero seguirán funcionando y la información disponible será consistente.

    - A y P: durante un fallo de uno de los nodos (o falta de comunicación) la información estará disponible pero puede que no sea consistente.

Mediante estas parejas se pueden clasificar los sistemas de almacenamiento de datos y conocer a que categoría pertenece cada uno, y cuáles son las características que pueden proporcionar y así adaptarlas a las necesidades existentes.

Figura4 Bases de datos NoSQL a las que pertenecen según el Teorema CAP

Modelo de Explotación

El modelo de explotación seleccionado para la obtención de información de Tweets fue C y A, debido a que en caso de no contar con una conexión a la red social de twitter, se pueda trabajar con información de interés recolectada en servidores locales, considerando como la opción más viable la extracción de información utilizando el modelo de análisis de sentimientos, habiéndose dividido el proyecto en las siguientes 5 etapas.

- Descarga de Tweets
- Procesamiento de Tweets (Obtención de Metadatos)
- Análisis de sentimientos
- Almacenamiento
- Desarrollo del servidor

Análisis de sentimientos.

El análisis de sentimientos, también llamado minería de opiniones, consiste en la extracción de información emitida por un usuario (post, blogs, etc) para su estudio y clasificación [9].

Proceso del Análisis de Sentimientos [10]

1. Tokenización o separación de palabras
2. Corrección de palabras
3. Etiquetación gramatical
4. Identificación y categorización de palabras
5. Análisis morfológico de la oración
6. Obtención de propiedades

Propiedades de un mensaje [9]:

Polaridad: Identificación y clasificación de sentimientos o emociones.

ISSN-2410-4000

- Polaridad positiva: Sensaciones de bienestar. Alegría, gratitud, esperanza.
- Polaridad neutra: No hay emociones.
- Polaridad negativa: Expresiones de malestar por parte del usuario. Enojo, tristeza

Intensidad de la emoción.

Fuerte

- Débil

Subjetividad:

- Objetivo: Es todo lo relativo a un objeto, sin importar lo que piense un sujeto.
- Subjetivo: Es todo lo relativo al sentir y pensar de un sujeto con respecto a un objeto.

La polaridad es la propiedad que describe el sentimiento o emoción con que fue escrita una oración, y es considerada la propiedad más importante del análisis de sentimientos, como se sabe un texto está formado de palabras y cada una por si sola tiene una carga polar, por ejemplo: La palabra amor tiene una carga positiva muy fuerte, por el contrario la palabra odio tiene una gran fuerza de polaridad negativa, de igual manera existen palabras neutras y palabras que anulan o invierten el significado de un mensaje. A continuación se explica de una forma más detallada este criterio[17].

Clasificación de palabras por polaridad:

- Palabras Positivas: Son las que expresan felicidad o gratitud, por lo regular son verbos y adjetivos.

  Ejemplos

  - Verbos: Amar, adorar, gustar, disfrutar, etc
  - Adjetivos: Hermoso, bonito, bueno, fascinante, etc

- Palabras Negativas: Son las que expresan enojo o tristeza, por lo regular son verbos y adjetivos.

  -

  Ejemplos

  - Verbos: Odiar, fastidiar, disgustar, enojar, etc
  - Adjetivos: Feo, horrible, malo, aburrido, etc

- Palabras Neutras: Son las que no representan emociones positivas ni negativas, por lo regular son sustantivos, artículos y pronombres.

  Ejemplos:

  - Sustantivos : Escuela, casa, perro, futbol, etc
  - Artículos: El, la, los, etc
  - Pronombres: Yo, tu, ellos, etc

- Negadores, inversores y/o anuladores: Palabras que al estar en una oración, cambian el significado de la misma y pueden ser adverbios de negación o nexos adversativos[11].

  Ejemplos:

  - Adverbios de negación : No, nunca, jamás, tampoco, etc [12]
  - Nexos adversativos: pero, sin embargo, por el contrario, no obstante, etc[13].

La obtención de la polaridad de cada palabra es la base para el análisis de sentimientos, sin embargo no es suficiente para la obtención de la polaridad de un texto.

Debido a la existencia de palabras que pueden invertir el significado de un texto, se desarrolló un modelo de prueba para evaluar las diferentes formas de expresar la polaridad (Ver Tabla 1)

Significado de abreviaturas en la Tabla 1

- P. Pos.: Palabras con carga positiva
- P. Neg.: Palabras con carga negativa
- P. Neu.: Palabras con carga neutral
- N.I.A.: Palabras que niegan, invierten y/o anulan el significado de una oración.

(Resultado esperado de la evaluación del texto)	Tipos de palabras que lo componen	Ejemplo de texto utilizado	Clasificación de polaridad de las palabras que componen el ejemplo			
			P. Pos	P. Neg.	P. Neu.	N.I.A.
Positivo simple	P. Pos. + P. Neu.	Ayer fue un excelente día	Excelente		Ayer, fue, un, día	
Positivo (Uso de NIA)	P. Neg. + NIA + P. Neu.	No estoy enojado		Enojado	Estoy	No
Negativo simple	P. Neg. + P. Neu.	Odio ir a la escuela		Odio	Ir, a, la, escuela	
Negativo (Uso de NIA)	P. Pos. + N.I.A. + P. Neu.	No me siento bien	Bien		me, siento	No
Neutro Simple	P. Neu.	Prueba 1			Prueba, 1	
Neutro (Complejo)	P. Pos. + P. Neg. + P. Neu. + NIA.	No estoy triste pero tampoco estoy feliz	Feliz	Triste	Estoy, estoy	No, pero, tampoco

Tabla 1 Descripción de SET de pruebas definido

A continuación se mencionan los algoritmos analizados y los resultados de la evaluación obtenidos. Una importante característica es que la implementación del sistema no consume muchos recursos y puede funcionar en Windows y Linux.

Principales algoritmos de análisis de sentimientos

NLTK

Es una plataforma desarrollada en Python para el (Procesamiento de lenguaje atural) PLN, proporciona interfaces fáciles de usar para más de 50 cuerpos y recursos léxicos, como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto para: Clasificación, tokenización, derivado, etiquetado, análisis y razonamiento semántico, NLTK solo está disponible para el idioma inglés.

Diccionario Marcado con Emociones y Ponderado para el Español

Proyecto liderado por el Dr. Grigori Sidorov [16]., consiste en un listado de palabras donde a cada una, se le atribuye una emoción y un porcentaje de probabilidad de ser usada al representar la emoción con la que se le relaciona. Es un trabajo enfocado al español y desarrollado en el IPN tiene contexto para el español de México, sin embargo no cuenta en la actualidad con análisis morfológico.

Sentiment140

Paquete de análisis de sentimientos en texto, creado por estudiantes de la Computación en la Universidad de Stanford, Funciona mediante clasificadores construidos a partir de algoritmos de aprendizaje automático. Soporta inglés y español[17] .

ISSN-2410-4000

TextBlob

TextBlob es una biblioteca de Python para el procesamiento de datos textuales. Proporciona una API simple para el procesamiento de lenguaje natural, realizando tareas comunes como el etiquetado, el análisis de sentimientos, la clasificación, traducción entre otras opciónes [18].

Sentiment Analysis of Meaning Cloud

Proporciona el servicio de API´s a desarrolladores y cuenta con su solución para el análisis de sentimientos. Como formato de salida ocupa JSON y obtiene características igual de importantes que la polaridad como la concordancia, la subjetividad, la confianza, y la ironía [19].

Bitext´s API

Utiliza Análisis lingüístico profundo basado en las gramáticas, lo que permite el análisis de la opinión no sólo en el nivel de la oración, sino también a nivel de la frase dentro de la oración. Esto es posible debido a que el análisis sintáctico identifica las diferentes frases (sintagmas nominales, frases adjetivas, frases verbales, etc.) y sus dependencias [20].

En la Tabla 2 se muestra la comparación de recursos para el análisis de sentimientos, a través de las características mas importantes de los recursos disponibles para llevar a cabo el análisis de sentimientos del proyecto.

Recurso y caracterís-tica	NLTK	Diccionario con Emociones	Sentiment 140	TextBlob	Senti-ment Analy-sis of Mea-ning Cloud	Bitext´s API
Tipo de recurso	Libre- ría	Diccionario	Librería y API	Librería y API	API	API
Idioma	Ingles	Español	Inglés y español	Ingles	Español	Inglés y Espa-ñol
Compatibilidad con Python 2.7	No	No aplica	Si	Si	Si	Si

Tabla 2 Tabla comparativa de los recursos para el análisis de sentimientos

Con referencia a la Tabla 2. Los recursos seleccionados para describir más a detalle y comprobar su eficacia fueron Sentiment140, Sentiment Analysis of Meaning Cloud y Bitext´s API, debido a que soportan su interacción con el lenguaje Python y están enfocados para el idioma español.

Algoritmos seleccionados evaluados

Sentiment140

Funcionamiento

1. El mensaje entra al proceso mediante una cadena JSON
2. Se conecta al servidor de sentiment140
3. Imprime en formato JSON el mensaje y su polaridad.

Parámetros de polaridad

- Positivo: 4
- Neutro: 2
- Negativo: 0

Código

```
import urllib2
import json
s = '{"data": [{"text": "Mensaje 1 "},{"text":
"Mensaje 2"}]}
response =
urllib2.urlopen('http://www.sentiment140.com/a
pi/bulkClassifyJson', s) # request to server
page = response.read() # get the response
print page # print the result
json.loads(page) # parse the result. The result is
in JSON format
```

Sentiment Analysis of Meaning Cloud

Funcionamiento

1. Se importan las librerías httplib, urllib y JSON

2. Se declaran las variables: host, api, key txt y el modelo
3. Se realiza el proceso de análisis de sentimiento
4. Si las palabras del mensaje coinciden con el diccionario de la API, se creará la variable score_tag que indicará la polaridad del mensaje, en caso contrario se entiende que el mensaje no tenía palabras significativas y por ende la polaridad es neutra.

Parámetros de polaridad

- Positivo: P, P+
- Neutro: Neu
- Negativo: N, N+

Código

```
# -*- encoding: utf-8 -*-

import httplib
import urllib
import json

# We define the variables need to call the API
host = 'api.meaningcloud.com'
api = '/sentiment-1.2.php'
key = '97457e61bf6d326d29b3369464255000'
txt = 'Mensaje'
model = 'es-general' #// es-general/en-general/fr-general

# Auxiliary function to make a post request
def sendPost():
  params = urllib.urlencode({'key': key,'model': model, 'txt': txt, 'src': 'sdk-python-1.2'}) # management internal parameter
  headers = {"Content-type": "application/x-www-form-urlencoded","Accept": "text/plain"}
  conn = httplib.HTTPConnection(host)
  conn.request("POST", api, params, headers)
  response = conn.getresponse()
  return response

# We make the request and parse the response
response_text = sendPost()
data = response_text.read()
r = json.loads(data)

# Show the response
print "Response"
print "================"
print data
print "\n"

# Prints the global sentiment values
print "Sentiment: "
print "==========="

if data.find("score_tag") == -1:
  r = json.loads(json.dumps({"score_tag": "NEU"}))
  print r
  print 'Global sentiment: ' + r['score_tag']

else:
  r = json.loads(data)
  print 'Global sentiment: ' + r['score_tag']
```

Bitext´s API

Funcionamiento

1. Se importan las librerías httplib, urllib.
2. Se declaran las variables: usr (usuario), pwd (contraseña), el txt (mensaje), el id, el idioma.
3. Se declaran los parámetros de conexión y salida de información.
4. Se declaran los encabezados y el servidor.
5. Se realiza la conexión donde se lleva a cabo el análisis de sentimientos, y una vez que obtiene el resultado, cierra la conexión.
6. Se imprimen los datos en JSON.

Parámetros de polaridad

- Positivo: (0.0 : 8.0]
- Neutro: 0.0
- Negativo: [-8.0 : 0.0)

ISSN-2410-4000

Código

```
# -*- coding: utf-8 -*-
import httplib, urllib

usr = 'xxxxx'
pwd = 'XXXXX'

txt = 'mensaje'
id = '0001'

lang='ESP'

params = urllib.urlencode({'User': '%s' % usr,
'Pass': '%s' %pwd, 'Lang': '%s' % lang, 'ID': '%s'
%id, 'Text': '%s' % txt, 'Detail': 'Detailed',
'OutFormat':'JSON', 'Normalized': 'No', 'Theme':
'Gen'})

headers = {"Content-type": "application/x-www-
form-urlencoded"}

server="svc8.bitext.com"

conn = httplib.HTTPConnection("%s" % server)
conn.request("POST", "%s" % service, params,
headers)
response = conn.getresponse()
data = response.read()
conn.close()
print data
```

Resultados

Desempeño de algoritmos seleccionados

Se muestran los resultados de un set de prueba realizado para comprobar la eficacia de las APIs, en diferentes circunstancias.

Composición del set de pruebas

- 6 oraciones en español, cada una representando un tipo de polaridad (Tabla 1)
- 3 tweets tomados aleatoriamente, cada uno representando una polaridad diferente

Ejemplos propuestos				
Tipo de oración	Oración	Resultado		Acierto
		Esperado	Obtenido	
Positivo simple	Ayer fue un excelente día	4	2	0.5
Positivo (Uso de NIA)	No estoy enojado	4	2	0.5
Negativo simple	Odio ir a la escuela	0	2	X
Negativo (Uso de NIA)	No me siento bien	0	0	✓
Neutro simple	Prueba 1	2	2	✓
Neutro (Complejo)	No estoy triste pero tampoco estoy feliz	2	2	✓
			Eficacia	58.33 %
Tweets Aleatorios				
Positivo	La pregunta es si me gusta mi trabajo? La respuesta es NOOO... "ME ENCANTA" #ITESM #cuernavaca…	4	0	X
Negativo	Como para provocar ataques epilépticos! #BibliotecaTec #ITESMQro #odioElNuevoTec	0	2	X
Neutral	Conoce nuestra carrera en Ingeniería Civil #ITESM	2	2	✓
			Eficacia	33.33 %

Tabla 3 Prueba de algoritmo Sentiment140

Ejemplos propuestos				
Tipo de oración	Oración	Resultado		Acierto
		Esperado	Obtenido	
Positivo simple	Ayer fue un excelente día	P, P+	P+	✓
Positivo (Uso de NIA)	No estoy enojado	P, P+	P	✓
Negativo simple	Odio ir a la escuela	N, N+	N	✓
Negativo (Uso de NIA)	No me siento bien	N, N+	N	✓
Neutro simple	Prueba 1	Neu	Neu	✓
Neutro (Complejo)	No estoy triste pero tampoco estoy feliz	Neu	Neu	✓
			Eficacia	100%
Tweets Aleatorios				
Positivo	La pregunta es si me gusta mi trabajo? La respuesta es NOOO... "ME ENCANTA" #ITESM #cuernavaca…	P, P+	P+	✓
Negativo	Como para provocar ataques epilépticos! #BibliotecaTec #ITESMQro #odioElNuevoTec	N, N+	N	✓
Neutral	Conoce nuestra carrera en Ingeniería Civil #ITESM	Neu	Neu	✓
			Eficacia	100%

Tabla 4 Prueba de Sentiment Anaysis of Meaning Cloud

Ejemplos propuestos				
Tipo de oración	Oración	Resultado		Acierto
		Esperado	Obtenido	
Positivo simple	Ayer fue un excelente día	(0.0 : 8.0]	5.0	✓
Positivo (Uso de NIA)	No estoy enojado	(0.0 : 8.0]	2.0	✓
Negativo simple	Odio ir a la escuela	[-8.0 : 0.0)	-3.0	✓
Negativo (Uso de NIA)	No me siento bien	[-8.0 : 0.0)	-2.0	✓
Neutro simple	Prueba 1	0.0	0.0	✓
Neutro (Complejo)	No estoy triste pero tampoco estoy feliz	0.0	-1	0.5
			Eficacia	91.66 %
Tweets Aleatorios				

Positivo	La pregunta es si me gusta mi trabajo? La respuesta es NOOO... "ME ENCANTA" #ITESM #cuernavaca…	(0.0 : 8.0]	2.0	✓
Negativo	Como para provocar ataques epilépticos! #BibliotecaTec #ITESMQro #odioElNuevoTec	[-8.0 : 0.0)	-2.0	✓
Neutral	Conoce nuestra carrera en Ingeniería Civil #ITESM	0.0	0.0	✓
			Eficacia	100%

Tabla 5 Prueba de Bitext´s API

Comparación de desempeño entre algoritmos

Con los resultados obtenidos de las tablas 3, 4, y 5, se realiza la gráfica de comparación entre las APIs

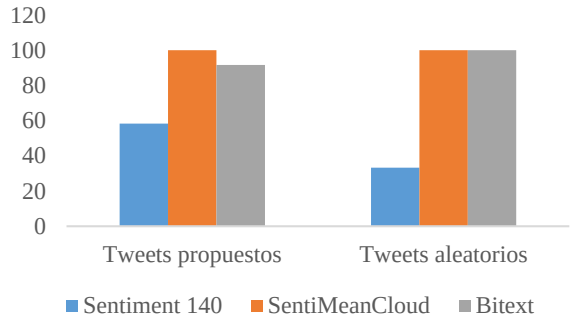


Gráfico 1 Comparacion de eficacia entre algoritmos

El recurso seleccionado es la API Sentiment Analysis of Meaning Cloud, debido a que su desempeño mostrado con el set de pruebas establecido fue inmejorable, la segunda opción en calidad de resultados fue la API Bitext´s utilizada, siendo la segunda alternativa recomendada para este tipo de análisis.

Aunque el API Sentiment Analysis of Meaning Cloud tenga 100% de eficacia, según la prueba que se realizó en este proyecto, no asegura que sea igual de eficaz cuando se ejecute en el sistema.

Arquitectura de Procesamiento seleccionada

Se pretende hacer un análisis en tiempo real que produzca resultados en forma gráfica a fin de facilitar la toma de decisiones a usuarios interactivos, los componentes de la arquitectura pueden ser vistos en la siguiente figura:

Figura 4 Arquitectura de análisis de tweets utilizada

Componentes de la arquitectura

1. Los usuarios publican en Twitter.
2. Un programa extrae y procesa Tweets de algún tópico o palabra con APIs existentes.
3. Se almacena en una base de datos NoSQL.
4. Los tweets se visualizan en un servidor web.

Pseudocódigo del programa que extrae y procesa los Tweets

1-. Leer tokens y keys.
2-. Autentificarse y conectarse a la API de

Twitter.

3-. Declarar la base de datos en MongoDB.
4-. Declarar la API de Análisis de Sentimientos

ISSN-2410-4000

Mientras Conexión exista:

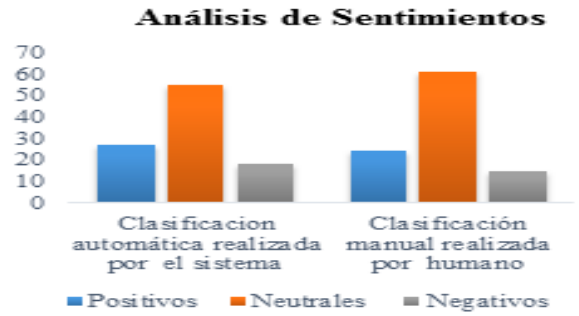
5.- Buscar palabras referentes.
6.- Obtener la Polaridad del tweet.
7.- Crear Arreglo JSON.
8.- Obtener los Metadatos e insertarlos al arreglo JSON.
9.- Insertar la Polaridad a Arreglo JSON.
10.- Envíar arreglo JSON a la BD.

Eficiencia de la aplicación al realizar análisis de sentimientos

Se reportan los resultados obtenidos de una muestra de 100 Tweets que contienen la palabra "Pachuca", de estos el sistema clasificó correctamente el 85% de los mensajes.

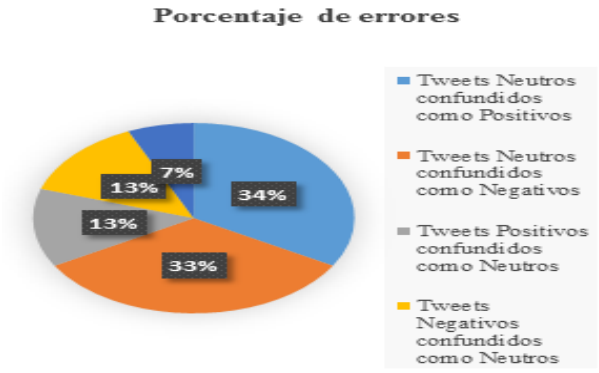
Gráfica 2 Eficacia en clasificación de Tweets

En la gráfica 3 se muestran los resultados de clasificación obtenidos por el sistema de análisis de sentimientos y el realizado por un experto humano. Se puede observar que en ambos casos mas del 50% son Tweets Neutrales.



Gráfica 3 Resultados de clasificación utilizando Análisis de Sentimientos

Un análisis más detallado de los tweets incorrectamente clasificados, es mostrado en la siguiente gráfica.



Gráfica 4 Porcentaje de errores

Conclusiones

- Funcionamiento en diferentes sistemas operativos (Windows y Linux), el desarrollo fue probado en Windows 7 y Ubuntu 14 LTS, encontrándose tiempos similares de ejecución para las pruebas realizadas.

- Sistema personalizable para clasificar tweets con diferentes textos.

- Almacenaje clasificado de tweets en base de datos NoSQL (MongoDB), con el propósito de mejorar su procesamiento y recuperación.

- Fácil habilitación de una interfase de explotación con los datos clasificados.

- Posibilidad de clasificar los datos directamente de la red de twitter utilizando APIs existentes de algoritmos de twitter implementados en lenguaje Python.

ISSN-2410-4000

- Para la explotación de Tweets, no fue necesario realizar el procesamiento en paralelo de mapeo y reducción propuesto por el Big data, debido a que los mensajes son pequeños (140 caracteres) y el número de tweets obtenidos por la consulta fueron pocos para los textos de consulta seleccionados, sin embargo para el procesamiento de un mayor número de datos con diferentes frases el big data consideramos que deberá ser utilizado.

- Las principales causas de error en la clasificación automática fueron:

  - Redacción de Tweets ambiguos o con ideas incompletas.

  - Existencia de gran cantidad reportes informativos que pueden ser interpretados con subjetividad, por ejemplo el clima o el tráfico.

  - Comparación entre entidades donde el tópico a analizar, es considerado como de poca relevancia.

  - Uso del sarcasmo.

- El uso de emoticons por parte de los usuarios y la capacidad de descifrado por parte de la API utilizada, ayuda a realizar un mejor análisis de sentimientos.

- Es importante señalar que para estudios más profundos de análisis de sentimientos, se requiere la colaboración de especialistas como: psicólogos, pedagogos, lingüistas, historiadores y sociólogos entre otros, que por su experiencia aporten puntos de vista importantes que complementen estos procesos.

- El estudio del análisis de sentimientos y sus aplicaciones, tiene muchas posibilidades de aplicación en el desarrollo social, económico y político.

- Se continuará con el desarrollo de un algoritmo propio que permita realizar el análisis de sentimientos de bases de datos almacenadas en manejadores NoSQL.

Referencias

1 Barranco R. (2012). ¿Qué es Big Data?. 15 de septiembre 2014, de IBM Sitio web:

http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

2 Bao, R., Flores, J., & González, F. (2009): Las organizaciones virtuales y la evolución de la Web. Lima, Perú. Universidad de San Martín de Porres, Fondo Editorial.

3 (2015). Aplicación de Tráfico en Internet. 27 de mayo del 2015, de © Alexa Internet, Inc.

Sitio web: http://www.alexa.com/topsites

4 (2015) What is Facebook?. 31 de Agosto de 2015, de Goodwill Community Foundation, Inc

Sitio web:
http://www.gcflearnfree.org/facebook101/2

5 Steve Thornton. (2009). Twitter versus Facebook: Should you Choose One?. 31 de agosto de 2015, de twitip

Sitio web: http://www.twitip.com/twitter-versus-facebook/

6 (2015). Que es LinkedIn?. 31 de agosto de 2015, de .woow.marketing
Sitio web: http://www.woow.marketing/que-es-linkedin/

7 Jaramillo M.. (des). CONOZCA GOOGLE+, EL CUARTO INTENTO DE RED SOCIAL DE GOOGLE. 31 de agosto de 2015, de enter.co Sitio web: http://www.enter.co/cultura-digital/redes-sociales/conozca-google-el-cuarto-intento-de-red-social-de-google/

8 Alarcón J. (2014). Fundamentos de bases de datos NoSQL MongoDB. 3 de marzo de 2015, de www.campusmvp.es Sitio web: http://www.campusmvp.es/recursos/post/Fundamentos-de-bases-de-datos-NoSQL-

9 (2012). Análisis de sentimiento: capturando la emoción. 27 de mayo de 2015, de www.daedalus.es Sitio web: http://www.daedalus.es/blog/es/analisis-de-sentimiento-capturando-la-emocion/

10 Casado A. (2013). Sistema de extracción de entidades y análisis de opiniones en contenidos Web generados por usuarios. 22 de Febrero de 2015, de Universidad Autonoma de Madrid Sitio web: http://ir.ii.uam.es/~fdiez/TFGs/gestion/leidos/2013/201309093AlvaroJoseCasadoValverde.pdf

11 (2015) Doble negación: no vino nadie, no hice nada, no tengo ninguna - See more at: http://www.rae.es/consultas/doble-negacion-no-vino-nadie-no-hice-nada-no-tengo-ninguna#sthash.kBwe2Q3S.dpuf. 28 de Abril de 2015, de Real Academia Española Sitio web: http://www.rae.es/consultas/doble-negacion-no-vino-nadie-no-hice-nada-no-tengo-ninguna

12 (2010). Ejemplos de Adverbios de Negación. 14 de junio de 2015, de http://www.gramaticas.net/ Sitio web: http://www.gramaticas.net/2010/09/ejemplos-de-adverbios-de-negacion.html

13 (2011). http://www.gramaticas.net/. 14 de Junio de 2015, de http://www.gramaticas.net/ Sitio web: http://www.gramaticas.net/2011/10/ejemplos-de-nexos-adversativos.html

14 NLTK Project. (2015). Natural Language Toolki. 15 de Abril de 2015, de www.nltk.org Sitio web: http://www.nltk.org/

15 NLTK Project. (2015). Installing NLTK. 15 de Abril de 2015, de www.nltk.org Sitio web: http://www.nltk.org/install.html

16 Díaz I , Sidorov G, ySuárez-Guerra S. (2014). Creación y evaluacion de un diccionario marcado con emociones y ponderado para el español. 10 de Abril de 2015 .DOI 10.7764/onomazein.29.5.

17 Alec Go, Richa Bhayani, y Lei Huang.. (Desconocido). General Information. 15 de Abril de 2015, de sentiment140 Sitio web: http://help.sentiment140.com/home

18 (2014). TextBlob: Simplified Text Processing. 15 de Abril de 2015, de http://textblob.readthedocs.org/ Sitio web: http://textblob.readthedocs.org/en/dev/

19 Meaningcloud. (2015). What is Sentiment Analysis?. 16 de Abril de 2015, de /www.meaningcloud.com Sitio web: https://www.meaningcloud.com/developer/sentiment-analysis/doc

20 Bitext. Sentiment Analysis API Service. 17 de Abril de 2015, de www.bitext.com Sitio web: http://www.bitext.com/text-analysis-technology/text-analysis-cloud-services-api/sentiment-analysis/