ESCUELA POLITÉCNICA NACIONAL
Facultad de Ingeniería de Sistemas (FIS)
Carrera de Ingeniería en Sistemas Informáticos y de Computación
Inteligencia de Negocios
PROYECTO SEGUNDO BIMESTRE

TEMA: Diseño e Implementación de un modelo de clasificación de sentimientos utilizando machine learning

El análisis de sentimiento utiliza el procesamiento de lenguaje natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva; por ende, está relacionado con la sociología en cuanto a las emociones y sentimientos ya que éstos son a menudo parte del proceso de toma de decisiones de una persona.
En este sentido, se ha escogido realizar el análisis de opinión pública utilizando datos de Twitter de la ciudad de Quito.

DOCUMENTO ESCRITO

1. Objetivo General

Implementar e investigar el funcionamiento de un clasificador de sentimientos utilizando los algoritmos de aprendizaje vistos en clase y los datos recolectados de Twitter para identificar tendencias de opinión en la ciudad de Quito.

Objetivos específicos:

- Crear un clasificador de sentimiento en español utilizando datos extraídos de Twitter para minar opinión pública en la ciudad de Quito
- Identificar y seleccionar las herramientas necesarias para procesar y analizar datos en tiempo real provenientes de Twitter

Este trabajo final pretende promover las siguientes habilidades de los estudiantes:

- Investigación
- Autoaprendizaje
- Pensamiento crítico

Fases del Proyecto

Adquisición de datos	Para esta fase es necesario recolectar datos utilizando un cosechador de tweets entregado al inicio de este curso a través del aula virtual. En dicho cosechador, se establecieron las coordenadas necesarias para capturar tweets correspondientes a la ciudad de Quito. Las herramientas utilizadas para esta fase corresponden a un script codificado en Python y una base de datos noSQL(CouchDb)
Pre-procesamiento	Será necesario filtrar los tweets correspondientes a la ciudad de Quito, utilizando vistas en la base de datos, puesto que se identificó previamente

	que los tweets recolectados tienen información proveniente de otros países como Perú o están escritos en lenguajes diferentes al español.
Procesamiento	Una vez que se tengan los tweets únicamente de la ciudad de Quito, será necesario procesar su campo "texto" para poder determinar la opinión pública. El procesamiento de texto de un tweet generalmente involucra la remoción de caracteres especiales, links, tags, etc. ¿Qué sucede en el caso de los emoticones?
Análisis	Es necesario realizar un análisis sobre el texto de cada tweet, para minar la opinión pública de la ciudad de Quito. En este caso se deberá diseñar un clasificador de sentimientos en español de tal forma que se logre un nivel de clasificación alto. En este sentido será muy importante determinar las características necesarias (i.e vector de características) para lograr una alta precisión.
Presentación	Para esta fase será necesario analizar los tweets referentes a cada zona de la ciudad de Quito que permita identificar cuál es la tendencia. Para esta fase se sugiere la utilización de herramientas con las cual sea posible presentar gráficos, tablas, o elementos que permitan comprender fácilmente los resultados. En esta fase será importante contar con visualizaciones que permitan responder a las siguientes preguntas: • Cuáles son las horas del día en donde la gente se siente mas feliz • Cuál es la emoción, en general, de la gente de la ciudad de Quito

2. Máximo de palabras

El documento escrito debe contener un máximo de 2500 palabras más o menos el 5%. Se excluyen tablas, figuras, referencias y apéndices.

3. Presentación del documento

Contenido esperado del trabajo escrito

- **Introducción**
- **Método**
- **Resultados**
- **Conclusiones y trabajo futuro**

El documento escrito deberá ser entregado a través de la herramienta de detección de plagio TurnitIn utilizando el enlace en al aula virtual.

4. Código

El código deberá ser subido a un repositorio GitHub (incluir un enlace al mismo en el documento) con los comentarios respectivos de las partes más importantes del mismo.

Incluir también una captura de pantalla de la información contenida en su repositorio, a parte del enlace al mismo en GitHub. También es importante tomar en cuenta que dentro de cada repositorio del proyecto, se deberán incluir los nombres de los colaboradores y un archivo README.md que indique claramente de qué se trata el proyecto, instrucciones de instalación y funcionamiento.

5. Penalidades por entrega atrasada de trabajos

Con el fin de garantizar igualdad para todos los estudiantes, las tareas deben ser completadas dentro del plazo especificado. Las presentaciones tardías significarán un 20% de reducción en la nota por cada día retraso. Los trabajos entregados después de dos días de la fecha final, ya no serán receptados y tendrán una nota de cero.

Los trabajos que excedan los límites de palabras en un 5%, incluyendo notas al pie, tendrán una penalidad del 10% de la nota total.

Los trabajos que excedan el límite de palabras en un 25% o más tendrán una penalidad del 25% de la nota total.

6. Plagio

El material de otras fuentes presentación sin el pleno reconocimiento (plagio) está fuertemente penalizado. Las sanciones por plagio pueden incluir una calificación de cero en el trabajo o una calificación de cero al semestre dependiendo de la gravedad de la acción.

7. Criterios de Calificación

Criterios de calificación escrito	**Posible calificación**
Introducción, presentación, ortografía y gramática	5
Método	10
Resultados y análisis	5
Conclusiones	10
Total	**30**

ESCUELA POLITÉCNICA NACIONAL
Facultad de Ingeniería de Sistemas (FIS)
Carrera de Ingeniería en Sistemas Informáticos y de Computación
Inteligencia de Negocios
PROYECTO SEGUNDO BIMESTRE

PRESENTACIÓN

Cada pareja tendrá un promedio de 20 minutos para realizar la exposición de su solución.

En la presentación de debe evidenciar el funcionamiento de la solución y se realizarán las pruebas necesarias para verificar el resultado de la exactitud del clasificador.

Criterios de calificación presentación	Posible calificación
Explicación clara del diseño del sistema	5
Explicación clara de problemas y soluciones	5
Manejo de recursos[1]	5
Interacción con el público	5
Respuesta correcta a preguntas	5
Organización y estructura de la presentación.	5
Manejo apropiado del tiempo	5
Total	**35**

[1] http://www.eoi.es/wiki/index.php/El_manejo_de_recursos_en_la_presentaci%C3%B3n_en_Direcci%C3%B3n_de_personas_y_habilidades