MCDI

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO

GERENCIA DE CAPITAL HUMANO

POSGRADOS

DETECCIÓN DE SPAM DE EMAIL MEDIANTE APRENDIZAJE AUTOMÁTICO

Tesis
Que para obtener el grado de MAESTRO EN CIENCIAS DE DATOS E INFORMACIÓN

Presenta:

Alvaro Miguel Toriz Proenza

Asesor:

Daniel Alejandro Cervantes Cabrera

Ciudad de México, marzo, 2023

CONACYT

INFOTEC POSGRADOS

Autorización de impresión

Agradecimientos

Tabla de contenido

Índice de figuras

Índice de cuadros

Siglas y abreviaturas

(Sigla o abreviatura): Nombre completo

(sigla o abreviatura): Nombre completo

Glosario

latex Is a mark up language specially suited for scientific documents.

Introducción

El llamado spam es un tipo de mensajes de texto no deseado ni solicitado, que es recibido a través de alguna plataforma digital, principalmente en el correo electrónico, pero también puede llegar en mensajes SMS (Short Message Service) o en plataformas sociodigitales como Twitter, etc.

Debido a que este tipo de mensajes es de una inmensa variedad y a que puede saturar los servidores de correo electrónico, existe un gran cantidad de investigación y herramientas computacionales dedicados a su filtrado. Por ejemplo los filtros Bayesianos los cuales buscan ocurrencias de palabras particulares en los mensajes. Así, para una palabra particular w, la probabilidad de que aparezca un spam es calculada por el número de veces que aparece en un conjunto grande de correos de spam y el número de veces que aparece en un conjunto grande de correos no spam.

Este trabajo se enfoca en los mensajes de spam que llegan por email. Mediante el uso de técnicas de Aprendizaje de Máquina, también llamado Aprendizaje Computacional o Aprendizaje Automático se podría facilitar la correcta clasificación de mensajes, para detectar y detener los mensajes de spam. Dicho Aprendizaje Automático, puede ser Supervisado o No Supervisado.

El objetivo, en el Aprendizaje Supervisado, es que se aprenda una función que se aproxime mejor a los resultados deseados. Se tiene un conocimiento previo de cuáles deberían ser los valores de salida para las muestras. El Aprendizaje Supervisado se realiza normalmente en el contexto de la clasificación. En el Aprendizaje No Supervisado, el objetivo es inferir la estructura natural presente dentro de un conjunto de puntos de datos. Algunos casos de uso comunes son el análisis exploratorio y la reducción de la dimensionalidad. El Aprendizaje No Supervisado se realiza normalmente en el contexto de la agrupación.

Algoritmos de Aprendizaje Supervisado

Árboles de decisión: El objetivo es crear un modelo que prediga el valor de una variable

objetivo en función de varias variables de entrada.

Clasificación de Naïve Bayes: un modelo probabilístico que se utiliza para tareas de clasificación, basado en el teorema de Bayes.

Regresión por mínimos cuadrados: Utilizada para tratar de encontrar el límite de decisión óptimo Regresión Logística: Probabilidad de que la salida del modelo, entre 0 y 1, pertenezca a un determinado grupo o clase.

SVM: Máquinas de vectores de soporte (Support vector machine), separa las clases en dos espacios lo más amplio posibles, mediante un hiperplano de separación mediante un vector de soporte. Random forest,: consiste en una gran cantidad de árboles de decisión individuales que operan como un conjunto.

Redes neuronales: Simulando el funcionamiento de las redes neuronales, es posible utilizar algoritmos que combinan varios de los metodos mencionados, para hacer más eficiente la clasificación.

Algoritmos de Aprendizaje No Supervisado

Algoritmos de clustering: Trata de encontrar una estructura de agrupamiento en una colección de datos no etiquetados.

Análisis de componentes principales: Utiliza una transformación ortogonal para convertir un conjunto de observaciones de variables en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales.

Capítulo 1

Los Datos y el Problema

1 Capítulo 1

latex Hello world! [1]

Ejemplo de cita. [3]

Otro ejemplo de cita. [2]

1.1. Identificación de las fuentes de datos

Para la investigación, se ocupará el dataset Enron-Spam. El dataset Enron-Spam es un conjunto de correos electrónicos utilizado principalmente para el entrenamiento de algoritmos de filtro de spam y análisis de texto. La historia de este conjunto de datos se remonta al escándalo de contabilidad de la compañía Enron, una empresa de energía que colapsó en 2001.

Durante la investigación realizada por el Departamento de Justicia de los Estados Unidos, se recolectaron una gran cantidad de correos electrónicos de la empresa, incluidos correos electrónicos personales y de negocios. En total, el conjunto de datos Enron-Spam está compuesto por alrededor de 500,000 correos electrónicos, con una mezcla de correos legítimos y spam.

Las características del conjunto de datos Enron-Spam incluyen la gran cantidad de correos electrónicos disponibles, lo que lo hace una opción popular para la investigación de filtrado de spam y análisis de texto. Además, los correos electrónicos fueron escritos en un entorno empresarial, lo que los hace diferentes de los correos electrónicos personales y redes sociales.

Entre las ventajas del conjunto de datos Enron-Spam, destaca la posibilidad de entrenar algoritmos de filtrado de correo no deseado con una gran cantidad de datos y con correos electrónicos de una fuente real y variada. Además, estos correos electrónicos se escribieron usando una variedad de estilos de escritura y cubren una amplia gama de temas, lo que hace que sea un conjunto de datos desafiante e interesante para análisis.

Entre las desventajas del conjunto de datos Enron-Spam, destaca su tamaño, ya que al ser muy grande, puede ser difícil trabajar con él en algunas herramientas y plataformas. Asimismo, los correos electrónicos están marcados como legítimos o spam, pero no tienen etiquetas adicionales que los clasifiquen en términos de contenido o intenciones específicas.

Muchas investigaciones han utilizado el conjunto de datos Enron-Spam en sus estudios, principalmente en las áreas de aprendizaje automático, minería de datos y análisis de texto. Los resultados de estas investigaciones han permitido la mejora de los sistemas de filtrado de correo no deseado para proporcionar a los usuarios una experiencia más óptima en el uso del correo electrónico.

En síntesis, el conjunto de datos Enron-Spam es un conjunto de correo electrónico amplio, útil y desafiante para la investigación relacionada con el spam en correo electrónico y el análisis de texto. Además, la fuente real y legítima de los correos electrónicos lo convierte en una excelente opción para el entrenamiento de algoritmos de filtro de correo electrónico.

1.2. Planteanimiento del problema

El Aprendizaje Computacional o Machine Learning puede contribuir a detectar los mensajes de spam. Representa un verdadero reto, porque la variedad de los tipos de mensajes de este tipo es inmesa, pero la aplicación de aprendizaje profundo y otros métodos de IA (Inteligencia Artificial) puede contribuir a la correcta clasificación de este tipo de mensajes con una aceptación significativa. Se llevará a cabo un anális comparativo de algunos de los algoritmos que se han utilizado para lograr este objetivo.

1.3. Preguntas, hipótesis y objetivos

Objetivo General

Estudio e implantación de métodos de aprendizaje automático para la detección automática de correos electrónicos no deseados, con el objetivo de mejorar la detección en sistemas de correo de código abierto.

Objetivos Específicos

1. Estudio de la metodología de detección de SPAM utilizando diferentes métodos basados en aprendizaje automático y en particular aprendizaje profundo.

2. Implantación de los métodos descritos en el punto 1.

3. Validación y comparación de los métodos descritos en el punto 1.

4. Apoyo en el mejoramiento de detección de SPAM en correo institucional INFOTEC.

1.4. Creación de la base de datos

La base de datos es un conjunto de correos electrónicos clasificados como spam o ham (no spam). Se trata de 17,171 mensajes de spam y 16,545 mensajes de ham. El tipo de datos es texto, consistente en el asunto y el cuerpo del mensaje de correos electrónicos en inglés.

Capítulo 2

El Abordaje del Problema

2 Capítulo 2

2.1. Límites y alcances

Debido a que se pretende medir la eficiencia de los algoritmos de Aprendizaje Automático, se utilizarán las siguientes métricas: Exactitud. Tasas de error. Precisión. F1-Score.

2.2. Justificación

2.3. Limpieza y procesamiento de la base de datos

2.4. Análisis exploratorio de los datos

2.5. Ajuste de los parámetros de los modelos

Entre los algoritmos que se han utilizado para la detección de spam por email están: Filtrado Bayesiano. SVM (Support Vector Machine). Clasificador kNN. Red Neuronal. Clasificador AdaBoost. Algunos enfoques novedosos que se han propuesto para este fin: Aprendizaje Profundo. Redes Generativas Antagónicas.

Capítulo 3

Los Marcos y los Resultados

3 Capítulo 3

3.1. Marco teórico

3.2. Marco metodológico

3.3. Análisis estadístico de los resultados

3.4. Conclusiones

3.5. Estructura y formatos

Conclusiones

Conclusiones

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Bibliografía

[1] CARPIZO, J., AND VALADÉS, D. *El voto de los mexicanos en el extranjero.* UNAM, México, 1998.

[2] CASTRO MEDINA, A. L. *et al,Accidentes de tránsito terrestre. Estudios sobre el peritaje.* Porrúa-UNAM, México, 1998.

[3] KELSEN, H. *La teoría pura del derecho,* 3 ed. UNAM, México, 1969. trad. de Eduardo García Máynez.

ANEXOS

A Anexo 1

Índice alfabético