Data statements para el dataset de la mentoría “Ciencia de datos aplicada al diálogo entre docentes y alumnos ”

Introducción

Sabemos que los datos juegan un papel fundamental en el aprendizaje automático. Cada modelo de aprendizaje automático se entrena y evalúa utilizando datos. Las características de estos conjuntos de datos (en inglés, *datasets*) cambiarán completamente el comportamiento de un modelo. Es poco probable que un modelo funcione bien si su contexto de uso no coincide con sus datasets de entrenamiento o evaluación, o si estos datasets reflejan **sesgos no deseados**.

Los data statements para datasets son una metodología que permiten caracterizar los datos y hacer explícitos posibles riesgos de su uso. Tienen el potencial de aumentar la **transparencia** y la **responsabilidad** de los *productos* basados en esos datos.
Si tenemos un data statement nos podemos preguntar si un sistema entrenado en ese dataset es o no apropiado para cómo se planea utilizar el sistema. Un data statement ayuda también a mejorar la **reproducibilidad** de un modelo al documentar cómo difieren distintas versiones del mismo dataset.

Están destinados a abordar las necesidades de dos grupos de partes interesadas clave: creadores de datasets y consumidores de datasets. Para los **creadores** de datasets, el objetivo principal es fomentar una reflexión cuidadosa sobre el proceso de creación, distribución y mantenimiento de un conjunto de datos, incluidas las suposiciones subyacentes, los riesgos o daños potenciales y las implicaciones del uso. Para los **consumidores** de datasets, el objetivo principal es garantizar que tengan la información que necesitan para tomar decisiones informadas sobre el uso de un conjunto de datos.

La forma de elaborar un data statement consiste en entrevistar al creador o responsable del dataset y realizar la serie de preguntas que se enumeran en este documento. En la siguiente sección se detallan los pasos a seguir.

Mentoría: Ciencia de datos aplicada al diálogo entre docentes alumnos y docentes

Metadatos

Nombre del dataset (si tiene nombre, si no asignarle uno):
Existen dos datasets principales, uno llamado *messages_preprocessed* con los diálogos entre tutores y estudiantes y otro llamado *datadump* con metadata. ambos con formato csv.
Fuente (incluir link a la fuente si está disponible y describirla):
La entrevistada ha firmado un acuerdo de confidencialidad por lo que no es posible brindar la fuente de los mismos.
Experto del dataset: Patricia A. Loto
Autores del data statement: Veronica Bornancini y Sandra Olariaga.
Otros que contribuyeron a este documento: -no aplica.
Versión (si está disponible) y fecha aproximada de creación: Se desconoce la fecha exacta de creación del dataset, se estima como fecha aproximada el año 2016.
Licencia bajo la que se distribuye el dataset: los dataset son privados, por los que no cuentan con licencia de distribución.

Resumen del data statement (largo mínimo 50 palabras y máximo 70 palabras):

Los textos incluidos fueron los diálogos entre tutores y alumnos transcurridos durante una sesión de tutoría y metadatos asociados a esos diálogos, los cuales incluyen datos acerca de los tutores y de la sesión, como edad, género y nacionalidad de los tutores. Además de la marca de tiempo de la sesión, el texto de las sesiones también incluía emojis y emoticones.
El principal objetivo era analizar dichos datos para evaluar la calidad del servicio brindado. El texto está en el idioma inglés, ya que los estudiantes que utilizan dicha aplicación son mayormente de EEUU.
Usos no deseados:
Creemos que no se debería utilizar por ejemplo, para excluir o no contratar tutores de determinados países o edades, o para favorecer la contratación de tutores de un determinado género.

1. Motivación para la creación del dataset

Las preguntas de esta sección están destinadas principalmente a alentar a los creadores de datasets a articular claramente sus razones para crear el conjunto de datos y promover la transparencia sobre los intereses de financiación.

1.1 **¿Con qué propósito se creó el conjunto de datos? ¿Se tenía una tarea específica en mente? ¿Conoce otros datasets parecidos, y, en ese caso, cuáles son las diferencias? Proporcione una descripción.**

Los datos que se trabajaron en la mentoría analizada fueron recolectados de una aplicación de mensajería instantánea desarrollada para brindar tutorías en matemática, química o física a estudiantes de nivel secundario. Dicho conjunto de datos se creó para diversos fines, pero la experta sólo está al tanto de uno, que es el análisis de la calidad del servicio que brinda la la app. Teniendo en cuenta lo anterior, podemos afirmar que claramente, se tenía una tarea específica en mente al momento de la creación del dataset.

La experta desconoce otros conjuntos de datos parecidos.

1.2 ¿Quién creó el conjunto de datos (por ejemplo, qué equipo, grupo de investigación) y en nombre de qué entidad (por ejemplo, empresa, institución, organización)?

Los datos fueron recolectados a través de la app de tutoría, y fueron almacenados por la empresa dueña de la aplicación. El nombre de la empresa se mantiene en el anonimato por privacidad.

1.3 ¿Quién financió la creación del conjunto de datos?
La experta supone que la empresa dueña de la aplicación.

2. Composición del dataset

Estas preguntas están destinadas a proporcionar a los consumidores del conjunto de datos la información que necesitan para tomar decisiones informadas sobre el uso del conjunto de datos para tareas específicas.

2.1 **¿Qué representan las instancias que componen el conjunto de datos (por ejemplo, documentos, fotos, personas, países)? ¿Hay varios tipos de instancias (por ejemplo, películas, usuarios y calificaciones; personas e interacciones entre ellas; nodos y conexiones)? Proporcione una descripción.**

En *messages_preprocessed* : Cada fila del dataset representa un turno de tutoría, es decir un mensaje entre tutor y estudiante, y a su vez un conj. de turnos conforman un diálogo de una determinada sesión.
En *datadump*: Cada fila del dataset representa una sesión y contiene fecha, duración de la sesión, tiempo de respuesta, edad, género, país y calificación del tutor, medio de pago de la aplicación, asignatura consultada en la tutoría, entre otros.
Hay un solo tipo de instancia.

2.2 ¿Cuántas instancias hay en total (y de cada tipo, si corresponde)?

En el dataset de mensajes preprocesados se computaron 17.987 diálogos, y en el cual se identificaron un total de 120 tutores y aprox. 7000 estudiantes.

2.3 ¿De qué datos consta cada instancia? ¿Datos "sin procesar" (por ejemplo, texto o imágenes sin procesar) o características? En cualquier caso, proporcione una descripción y algunos ejemplos.

Mensajes preprocesados: texto e identificadores (id)

- Ejemplo de identificadores: id de tutor y de estudiante.
- Ejemplos de texto: mensajes entre tutor y estudiante, el cual podía contener: texto, emoticons, emojis, fórmulas, palabras en otro idioma, por ejemplo, en español.

Datadump*:* texto, booleanos y datos numéricos.

- Ejemplo de texto: nacionalidad del tutor, área de conocimiento del tutor, etc.
- Ejemplo de datos booleanos: por ejemplo, el campo student_complained, que hace referencia a si el estudiante presentó una queja sobre el servicio o no.
- Ejemplo de datos numéricos: edad, puntaje asignado a la sesión por parte del estudiante, entre otros.

Ejemplo de un diálogo:

```
-student: Can we look at where to start here ?
-student: Ok . No prob .
-student: Thank you !
-tutor: Thanks ! You were really nice .
-student: 😃
-tutor: It was fun doing maths with you .
-student: You too
-tutor: Thanks for using *** ! Have  a good one ! :)
```

2.4 ¿Hay una etiqueta o anotación de supervisión asociada con cada instancia? Si es así, proporcione una descripción.

El campo que se utilizó como etiqueta fue el puntaje asignado por los estudiantes a la sesión, el cual podía contener valores del 1 al 5, donde el 1 es el puntaje más bajo y 5 el más alto.

2.5 ¿Cree que hay algún aspecto de los datos que no ha sido representado y que podría llegar a resultar relevante para alguna aplicación o por sus efectos en algún segmento de la población (p. ej., falta información de género)? De ser así, proporcione una descripción.

No se disponía de datos referentes a los estudiantes: edad del alumno, nivel académico, ciudad de residencia y género del alumno. Disponer de estos datos hubiera permitido caracterizarlos y segmentarlos para ofrecerle un servicio más personalizado.
Se podría llegar a inferir el género del estudiante a partir de su nombre, lo cual podría ser mal usado para realizar discriminaciones basadas en género.

2.6 ¿El conjunto de datos contiene datos que podrían considerarse confidenciales (por ejemplo, datos que están protegidos por privilegios legales o por la confidencialidad del médico-paciente, datos que incluyen el contenido de las comunicaciones no públicas de las personas)? Si es así, proporcione una descripción.

No había datos confidenciales en ninguno de los dos dataset pero si consideramos que los mensajes entre tutor y estudiante son contenido de comunicaciones no públicas quizás podrían considerarse confidenciales.
En el hipotético caso de que el estudiante pudiera estar en la instancia de examen y estar realizando consultas al tutor respecto de la resolución del mismo sería muy difícil para el tutor poder darse cuenta de esta situación y en el caso de que sospechara de este escenario podría decírselo al estudiante pero con el riesgo de que el estudiante niegue esta situación.

3. Recopilación del dataset

Estas preguntas están destinadas a identificar características de la recopilación del dataset para advertir a consumidores del conjunto de datos sobre posibles sesgos emergentes debido a decisiones de recolección.

3.1 ¿Qué mecanismos se utilizaron para recopilar los datos, elegir la muestra y curar los datos (por ejemplo, aparatos o sensores de hardware, curación humana manual o automática, programa de software)? Esta pregunta es particularmente importante para datasets que son demasiado grandes para ser inspeccionados a mano. Explicar la racionalidad de la muestra puede ayudar a los usuarios de datasets a entender a qué tipos de usos el dataset puede generalizar bien o mal.

La recopilación de datos fue realizada a través de una aplicación web como se especificó en puntos anteriores, cabe aclarar que al dataset *messages_preprocessed* ya se le había realizado un preprocesamiento antes de entregarlo a los estudiantes participantes de dicha mentoría, incluyendo el anonimizado parcial del mismo. Ya que, debido a la complejidad de aplicar esta tarea al lenguaje natural, los diálogos no fueron anonimizados.

3.2 ¿Quiénes participaron en el proceso de recopilación de datos (p. Ej., Estudiantes, trabajadores digitales, contratistas) y cómo se les compensó (p. Ej., ¿Cuánto se les pagó a los trabajadores)?

La empresa dueña de la aplicación web se encargó de la recopilación de datos pero se desconoce cuál fue el personal específico que participó en dicho proceso o si hubo algún tipo de compensación para los involucrados.

3.3 ¿Durante qué período de tiempo se recopilaron los datos?

Entre los años 2016 y 2017, de acuerdo a los timestamps registrados en el dataset *datadump, el cual fue descripto en el punto 2.3.*

3.4 ¿El conjunto de datos se relaciona con las personas? De lo contrario, puede omitir las preguntas de la sección 4.

Si, se relaciona con personas: estudiantes y tutores.

4. Datos de personas

4.1 ¿El conjunto de datos incluye datos demográficos (por ejemplo, por edad, sexo)? De ser así, enumere qué datos demográficos se incluyen.
El conjunto de datos sólo incluye datos demográficos, como edad, país y género, de los tutores.

4.2 ¿Es posible identificar individuos (es decir, una o más personas físicas), ya sea directa o indirectamente (es decir, en combinación con otros datos) del conjunto de datos? Si es así, describa cómo.

Se pueden identificar estudiantes y tutores a través de un ID del tutor y del estudiante.

4.3 ¿El conjunto de datos contiene datos que podrían considerarse sensibles de alguna manera (por ejemplo, datos que revelan orígenes raciales o étnicos, orientaciones sexuales, creencias religiosas, etc.)?

Dentro del conjunto de datos se cuenta con el país de acceso y el género de los tutores que puede ser considerado información sensible.

4.4 ¿Recopiló los datos de las personas en cuestión directamente o los obtuvo a través de terceros u otras fuentes (por ejemplo, sitios web)?

Fueron obtenidos a través de un tercero, a través de la persona a cargo de la mentoría, la Dra. Luciana Benotti.

4.5 ¿Se notificó a las personas en cuestión sobre la recopilación de datos? Si es así, describa cómo se envió la notificación.
No aplica.

5. **Usos**

Estas preguntas están destinadas a alentar a los usuarios del dataset a reflexionar sobre las tareas para las que el dataset debe y no debe usarse. Al resaltar explícitamente estas tareas, los consumidores de datasets pueden tomar decisiones mejor informadas, previendo riesgos o daños potenciales.

5.1 ¿Para qué tareas creen que se puede usar este conjunto de datos?¿Ya se ha utilizado el conjunto de datos para alguna tarea? Si es así, proporcione una descripción.

Durante la mentoría se ha utilizado el conjunto de datos para predecir la satisfacción de los estudiantes. También se podría utilizar para:

- identificar el mejor y el peor tutor según los puntajes asignados por los estudiantes.
- si contáramos con el país de procedencia de los estudiantes se podría determinar cuáles son los países que más usan la aplicación. Aunque actualmente esto no es posible, ya que no se dispone de datos referentes a los estudiantes.
- conocer si existe algún tipo de relación entre el puntaje obtenido por el tutor en el exámen de admisión y el puntaje asignado por los estudiantes.
- realizar análisis de sentimientos
- identificar o caracterizar grupos de estudiantes
- analizar cuáles son las franjas horarias de mayor conexión de los estudiantes y con eso prever una mayor cantidad de tutores disponibles.

5.2 ¿Existe un repositorio que se vincule a alguno o todos los artículos o sistemas que utilizan el conjunto de datos? Si es así, proporcione un enlace u otro punto de acceso.

No se dispone de tal repositorio.

5.3 ¿Para qué (otras) tareas se podría utilizar el conjunto de datos?

Se ha respondido en el punto 5.1

5.4 ¿Hay algo sobre la composición del conjunto de datos o la forma en que se recopiló y preprocesó / limpió / etiquetó que pueda afectar los usos futuros? Por ejemplo, ¿hay algo que un futuro usuario pueda necesitar saber para evitar usos que podrían resultar en un trato injusto de individuos o grupos (por ejemplo, estereotipos, problemas de calidad del servicio) u otros daños indeseables (por ejemplo, daños financieros, riesgos legales)? Si es así, proporcione una descripción.

Al ser posible asociar el puntaje obtenido por los tutores a la nacionalidad, edad o género de los mismos, esto podría generar que se evite contratar tutores de ciertos países o de una determinada edad.

5.5 ¿Hay tareas para las que no se debería utilizar el conjunto de datos? Si es así, proporcione una descripción.

Creemos que no se debería utilizar por ejemplo, para excluir o no contratar tutores de determinados países o edades, o para favorecer la contratación de tutores de un determinado género.

5.6 ¿Algún otro comentario?
No.