UN PROCESO ÁGIL PARA ANALÍTICA DE TEXTOS USANDO **MÁQUINAS DE APRENDIZAJE**

- *Rodolfo Canelón*
  Profesor, Ingeniería civil en computación e informática, Facultad de Ingeniería Universidad Central de Chile.
- *Samanta Luna*
  Profesor, Ingeniería civil en computación e informática, Facultad de Ingeniería Universidad Central de Chile.
- *Nicolás Ulloa*
  Profesor, Ingeniería civil en computación e informática, Facultad de Ingeniería Universidad Central de Chile.

El presente artículo mostrará un proceso ágil en el análisis de texto mediante la implementación de máquinas de aprendizaje, para ello tendrá como caso de estudio el análisis de sentimientos en ambientes educativos denominado MaTexEdu. En el mismo, se analiza y contextualiza el proceso hecho por el uso de técnicas de Analítica de textos que ayudan a la compresión, clasificación y análisis de textos.

El proceso de analítica de textos abarca un completo análisis científico para deducir patrones y tendencias que existen en los datos. Dentro del proceso aquí planteado se definen 3 disciplinas: recolección de fuentes, la ingeniería de rasgos y formulación del modelo. Así mismo, cada una de estas disciplinas cuenta con un conjunto de actividades y artefactos.

El modelo MaTexEdu aquí planteado, muestra un motor analítico, para identificación de sentimientos, mediante la ubicación de las características presentes en el texto objetivo de investigación. Se realiza modelado utilizando distintos enfoques, como tener en cuenta presencia, frecuencia de términos, analizar partes del comentario, usar palabras, frases de opinión y uso de negaciones.

El análisis se realiza en español e inglés, para efectos de comparar grado de acierto en cada idioma. Posteriormente, se clasifica el sentimiento asignándole una polaridad. Dentro del modelo plateado, se utilizará los conceptos de procesos de desarrollo, reusabilidad de activos de software, ingeniería del dominio, modelos de variabilidad y líneas de producción de software aplicados para definir una vía en la cual estas aplicaciones serán implementadas. Por lo que, estos requisitos de variabilidad y adaptabilidad del software representan un interesante desafío dentro de la ingeniería de software. de máquinas de aprendizaje y/o diccionarios léxicos. Así mismo, para este estudio en particular, se utilizan enfoques de máquinas de aprendizaje y diccionarios léxicos.

El aporte de este trabajo es presentar un método que permita el desarrollo completo y detallado de la analítica de textos. Es por ello, que la contribución de este articulo al área del desarrollo de la ingeniería artificial, basado en líneas de producción de software es proponer un proceso para la ingeniería del dominio con un enfoque de calidad que permita formalizar las especificaciones de los rasgos de las familias de máquinas de aprendizaje.

Palabras Clave: *Ciencia de datos, analítica de textos, máquinas de aprendizaje, procesos ágiles, ingeniería del dominio, líneas de producción de software, ingeniería del Software, arquitectura de software, calidad del software.*

Introducción

Ágil es un término general que se refiere a varias metodologías que se centran en ser iterativas y en obtener productos y características tangibles rápidamente. Es así, como las metodologías ágiles permiten dar soluciones tangibles. Así mismo, la ciencia de datos, también se ha beneficiado de tomar fragmentos de conceptos ágiles.

La metodología Agile permite a los científicos de datos la capacidad de planificar y priorizar al crear hojas de ruta basadas en requisitos y objetivos. Esto también permite que los equipos técnicos brinden a los interesados una visión general y una comprensión de los costos totales asociados con cada objetivo general. Por lo que, ágil no se trata solo de trabajar en el software y los modelos; también se trata de alinear a los científicos de datos con el resto de la organización. Por lo tanto, todo el proceso crea una mejor alineación entre los científicos de datos y las partes interesadas al crear líneas constantes de comunicación.
A veces, una desalineación puede interponerse entre ingenieros y científicos de datos cuando los científicos de datos siguen esperando implementaciones de modelos mientras los ingenieros se preguntan qué están haciendo los científicos con la investigación aplicada y el análisis de datos. En este escenario, el proceso agile cierra la brecha entre ambos equipos para crear caminos más claros entre sus objetivos. La razón es que las metodologías ágiles hacen frente a realidades impredecibles de generar aplicaciones de análisis útiles a partir de los datos en bruto a escala.

En contraste con el desarrollo de software, los proyectos de ciencia de datos no se pueden prescribir o diseñar desde el principio, ya que es difícil conocer de antemano las técnicas y métodos más efectivos para el proyecto. En general, cada proyecto de ciencia de datos requiere que sigan diferentes caminos y prueben diferentes técnicas. Por lo tanto, estos proyectos afectados a ser iterativos requieren el complemento ágil para proyectos de ciencia de datos.

Las empresas líderes, con la incorporación de las metodologías ágiles, están construyendo plataformas de aprendizaje automático que dividen los datos de capacitación para volver a capacitarlos e implementarlos en modelos a través de API.

Actualmente en nuestra sociedad se presenta mucha información de distintas e infinitas maneras, pensamientos, emociones, gustos, tendencias, noticias y cientos más, dentro de esta información surge una problemática inicial de cómo se deriva la información más relevante o mejor dicho información de alta calidad a partir del texto, creando así la analítica de texto. Por lo

que, la información de alta calidad se obtiene al idear patrones y tendencias por medios estadísticos. La analítica de texto implica un análisis y adiciones de algunas características lingüísticas derivadas y la eliminación de otras adhiriéndolas a una base de datos, derivando ciertos patrones y con finalidad de evaluación e interpretación de resultados. El análisis de texto implica una recuperación de información de alta calidad para estudiar una distribución de cierta frecuencia de palabras, reconocimientos de patrones, etiquetados, vínculos, asociaciones, visualización y un análisis predictivo.

La analítica de texto es el descubrimiento de información nueva, previamente desconocida, mediante la extracción automática de información de diferentes recursos escritos [1].

Es el proceso de transformar datos de texto no estructurados en información significativa y procesable. La analítica de texto utiliza diferentes tecnologías de inteligencia artificial para procesar datos automáticamente y generar información valiosa, lo que permite a las empresas tomar decisiones basadas en datos.
Para las empresas, la gran cantidad de datos que se generan todos los días representa tanto una oportunidad como un desafío. Los datos ayudan a las empresas a obtener información valiosa sobre las opiniones de las personas sobre un producto o servicio.

Se podría pensar en todas las ideas potenciales que podría obtener al analizar correos electrónicos, reseñas de productos, publicaciones en redes sociales, comentarios de los clientes, tickets de soporte, entre otros. Por otro lado, está el dilema de cómo procesar todos estos datos y allí es donde la analítica de texto juega un papel importante.

Las personas y las organizaciones generan toneladas de datos todos los días. Las estadísticas afirman que casi el 80% de los datos de texto existentes no están estructurados, lo que significa que no están organizados de una manera predefinida, no se pueden buscar y es casi imposible de administrar. En otras palabras, simplemente no es útil. Por lo que, ser capaz de organizar, categorizar y capturar información relevante a partir de datos sin procesar es una preocupación y un desafío importante para las empresas.

La Analítica de texto es fundamental para esta misión, en un contexto empresarial, los datos de texto no estructurados pueden incluir correos electrónicos, publicaciones en redes sociales, chats, tickets de soporte, encuestas, entre otros. La clasificación manual de todos estos tipos de información a menudo resulta en fallas. No solo porque requiere mucho tiempo y es costoso, sino también porque es inexacto e imposible de escalar. Desde lo cual, la analítica de texto viene demostrando ser una forma confiable y rentable de lograr precisión, escalabilidad y tiempos de respuesta rápidos

La analítica de texto ayuda a analizar grandes cantidades de datos sin procesar y a encontrar información relevante, la cual, combinada con el aprendizaje automático, puede crear modelos de análisis de texto que aprenden a clasificar o extraer información específica en función de la formación previa.

En tal sentido, la estructura del presente artículo, además de la introducción y las conclusiones, es la siguiente: Alcance y presentación, muestra la estructura del marco referencial MaTexEdu; Las disciplinas de recolección de fuentes, la ingeniería de rasgos y formulación del modelo junto a la definición de las actividades y sus artefactos finalmente, Conclusiones.

Alcance y Presentación

Diversas investigaciones confirman que la tecnología de aprendizaje automático es muy eficiente para predecir situaciones. Esta técnica se aplica a través del aprendizaje de datos anteriores [4], [5]. El modelo desarrollado en este trabajo utiliza técnicas de aprendizaje automático y crea una nueva forma de ingeniería de datos y selección de características, para medir el rendimiento del modelo. Así mismo, fue preparado y probado a través del entorno Python 3.8 y sus librerías: matplotlib, pandas, pandasgui, spacy, sklearn, numpy, textblob, json y utilizando api´s tweepy, facebook para las redes sociales respectivas y almacenamiento de datos en formatos csv, postgres con modelos date lake.

El proceso planteado denominado MaTexEdu, en cuya exposición se utilizará la notación SPEM 2 [2], muestra la estructura de solución al problema planteado. El mismo fue generado como un proceso que permita la generalización del Modelo Predictivo de análisis de sentimientos en diversos dominios y el objetivo final es obtener un modelo de Clasificador Base para la familia de aplicaciones análisis de sentimientos.

El símbolo denotado por identifica la actividad afectada en la disciplina y el artefacto generado. De manera que, se proponen y adaptan un conjunto de técnicas específicas para la definición de las principales actividades y artefactos [3]. A continuación, se presentan las disciplinas del proceso de Máquinas de aprendizaje MaTexEdu. En proceso general plateado para la minería de texto se distinguen 3 disciplinas principales: Recolección de fuente, Ingeniería de rasgos y formulación del modelo (ver figura 1).

Análisis de Fuentes

En la recolección de fuente para realizar este proceso, la presente propuesta comprende 4 actividades en su primera disciplina de recolección de fuentes y está compuesto por las actividades: recolección inicial de texto, descripción de texto, detección del lenguaje y verificación de calidad de los textos (ver figura 2).

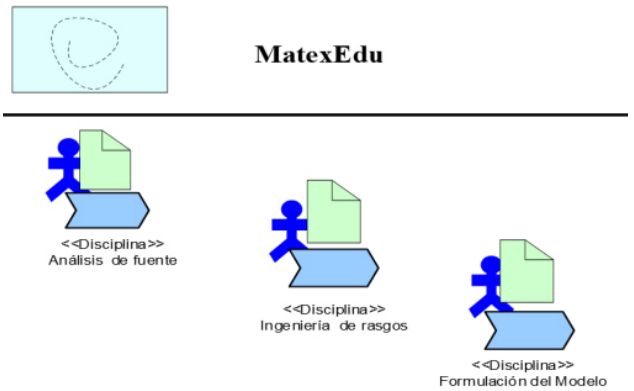

Figura 1

Disciplinas MaTexEdu

La fase de Recopilación inicial de texto es fundamental para el desarrollo del análisis del texto, ya que en esta etapa se precisa un conjunto de información que aún no ha sido procesada con el fin de obtener información de alta calidad. Al tener un conjunto de información la fase de descripción de texto no identifica la información como tal. En ella se deben recopilar los datos para su análisis. Por ejemplo, analizar los Twitter de la Universidad Central en su cuenta oficial de la Facultad de Ingeniería y Arquitectura.

El detector del lenguaje identifica el idioma de la información indicándonos el porcentaje de lenguaje con una tasa asociada. Por ejemplo, si el idioma es español, se identificará con un 100 la tasa española, verificando la calidad garantizada del texto. En la verificación de la calidad de los textos se Identifican las propiedades de calidad que deben ser garantizadas para su uso y con relación establecida.

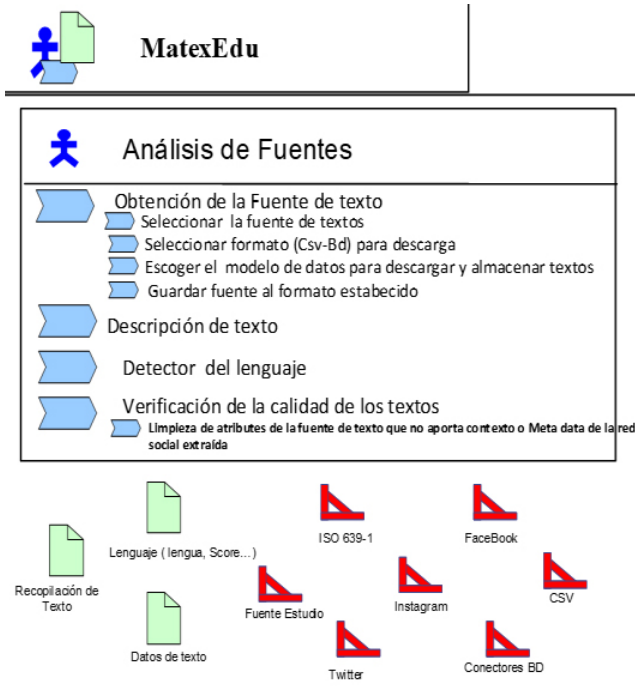

Figura 2

Análisis de fuentes, MaTexEdu

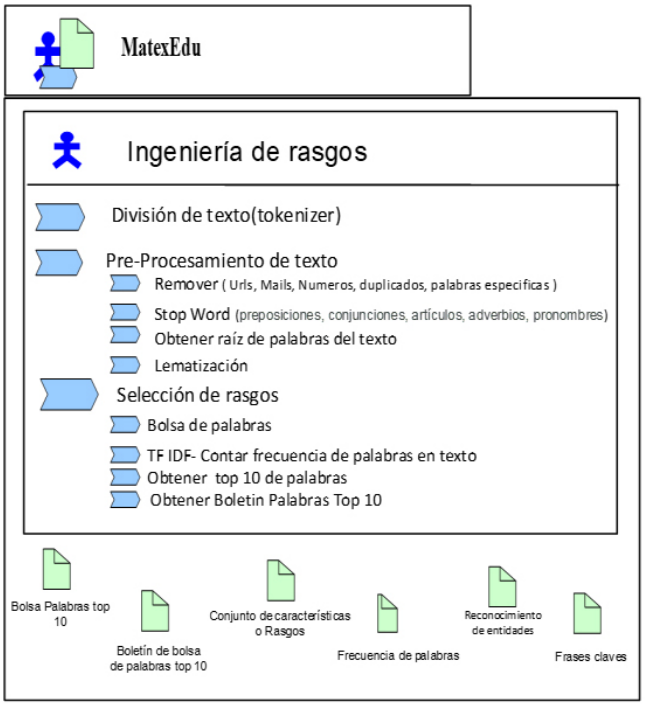

Figura 3

Ingeniería de Rasgos, MaTexEdu.

Ingenieria de rasgos

En el modelo MaTextEdu, la ingeniería de rasgos se especializa en todas las actividades que involucren las filtraciones del texto, los datos que deben ser limpiados y estructurados de tal manera que esta información sea de una mayor calidad para poder ser usada en las siguientes actividades, esta información será de utilidad para identificar, patrones, tendencias, pensamientos, entre otros. El preprocesamiento incluye el eliminar todas las partes del texto innecesarias como, por ejemplo, se pueden eliminar palabras como "Y", "pero", "es" y "la". Para poder extraer las características del texto, se requiere aplicar una serie de técnicas de preprocesamiento de texto, las cuales van a facilitar la tarea de la estructuración. Las fases principales de la disciplina de Ingeniería de rasgos constan de 3 actividades principales: División de texto, Preprocesamiento de texto y Selección de rasgos, mostrados en la figura 3.

Tokenization

Es la división de una cadena de caracteres en bruto en elementos de interés. A menudo, estos elementos son palabras, pero es posible que también se mantengan los números o puntuación.

Preprocesamiento de texto

Remove: Remueve palabras no innecesarias para la información de alta calidad. Remueve URLs, Mails, palabras duplicadas, números y palabras específicas.
Stop Words: eliminación de palabras vacías. Se trata de palabras que no aportan nada al análisis o estudio del texto, son por ejemplo conjunciones o preposiciones.

Inicialmente, el proceso busca reducir las palabras a un lema o raíz. La técnica de lematización es la derivación. Desde la cual, se obtiene el "lema", que es una palabra raíz. Finalmente, después de la lematización, se obtiene una palabra válida que significa lo mismo en el contexto.

Selección de rasgos

En este subproceso inicial, se utiliza el procesado del lenguaje para representar documentos ignorando el orden de las palabras. El artefacto "bolsa de palabras" (Bag of Words) donde cada documento parece una bolsa que contiene algunas palabras. Por lo tanto, este método permite un modelado de las palabras basado en diccionarios, donde cada bolsa contiene unas cuantas palabras del diccionario.

Así mismo, la frecuencia (TF-IDF) en la que se repite un determinado termino de dicho documento. Seguidamente, la incrustación de palabras (Word embedding) es el nombre de un conjunto de lenguajes de modelado y técnicas de aprendizaje en procesamiento del lenguaje natural (PLN) en dónde las palabras o frases del lenguaje natural son representadas como vectores de números reales.

Asimismo, está presente el método para obtener el top 10 o las más frecuentes palabras repetidas en bolsa de palabras. Finalmente, el boletín de palabras cuyo método es utilizado para generar las palabras frecuentemente usadas y generadas en su contexto de uso.

Formulación del Modelo

En sus actividades, el modelo MaTexEdu en su fase de análisis de sentimientos mostrada en la figura 4, usa la plataforma, de código abierto, para construir programas en Python con NLTK (Natural Language Toolkit) que trabajan con datos del lenguaje humano. En dicha plataforma se proveen librerías que permiten interfaces para más de 50 corpora y recursos léxicos. Las librerías permiten procesamiento de texto para clasificación, tokenización, stemming, tagging, parseo y razonamiento semántico, además de librerías robustas para NLP (Natural Language Processing).

Mediante estos elementos de software se permite categorizar textos, analizar estructuras lingüísticas y más. NLTK incorpora librerías para análisis de sentimientos en español, por lo que se pueden usar dos técnicas para el desarrollo de los análisis: usar aplicación para español o traducir texto a inglés y analizarlo con aplicación en inglés para estos propósitos. Por otro lado, se puede utilizar Vader [6] con NLTK, el cual es una herramienta de análisis para sentimientos, desarrollada y especialmente diseñada para analizar sentimientos expresados en redes sociales o textos de

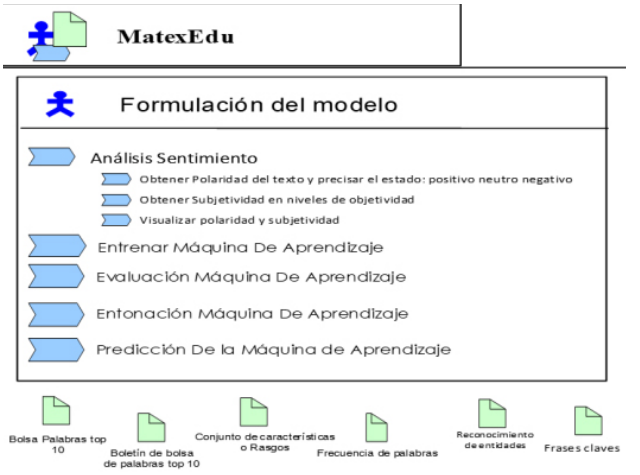

Figura 4

Formulación de modelos, MaTexEdu.

otros dominios. Por ejemplo, si las observaciones son palabras recopiladas en documentos, postula que cada documento es una mezcla de un pequeño número de temas y que la presencia de cada palabra es atribuible a uno de los temas del documento.

Enfoque de máquinas de aprendizaje

Mediante esta vía de análisis de texto de manera automática, en modo supervisado o no supervisada, la primera se basa en conjuntos de entrenamiento, los cuales serán usados para catalogar el resto de las opiniones encontradas en la fuente utilizada, realizando pruebas y luego validándolas. Las principales técnicas de este método son: Support Vector Machines (SVMs), Redes Neuronales, Naive Bayes, Redes Bayesianas, y clasificadores de Máxima Entropía, entre otras.

De esta manera, se utiliza la categoría gramatical de las palabras, la presencia y frecuencia de algunos términos y su composición semántica.

La mayoría de estos métodos, sin embargo, van acompañados de algún diccionario que entrega información a priori de los términos para obtener las polaridades respectivas. En algunos casos, estos diccionarios son realizados por personas y en otros se ocupa un sistema semiautomático.

Entrenar Máquina de Aprendizaje

Para entrenar el modelo de la máquina de aprendizaje, se debe especificar la fuente de datos de entrenamiento de entrada, nombre del atributo de datos que contiene el objetivo a predecir, instrucciones de transformación de datos requeridas y parámetros de entrenamiento para controlar el algoritmo de aprendizaje. Durante el proceso planteado, en la disciplina previa se selecciona el algoritmo de aprendizaje correcto según el tipo de objetivo que se especifica en la fuente de datos de entrenamiento.

Evaluación Máquina de Aprendizaje

Para medir el rendimiento del modelo propuesto se observa la tasa de error cometido por el mismo, es decir, se calcula el porcentaje de casos clasificados de forma incorrecta sobre el conjunto de datos que componen la muestra total.
Se procede a utilizar los grupos de datos, generadas en las etapas previas, las muestras de entrenamiento y predicción. Por lo que, los datos de entrenamiento son utilizados para descubrir relaciones potencialmente predictivas y patrones escondidos en los datos, mientras que los datos de predicción serán utilizados para evaluar la potencia y la utilidad de la predicción.

Entonación Máquina de Aprendizaje

Esta actividad permite ajustar los parámetros del modelo para mejorar las predicciones y precisión, buscando obtener la tasa de error esperada. Por lo que, se plantea retornar a la fase de entrenamiento haciendo antes una nueva configuración de parámetros del modelo. Así mismo, permite desarrollar acciones para tal fin, por ejemplo:

- Incrementar la cantidad de veces que iteramos los datos de entrenamiento.
- Calcular la tasa de aprendizaje y hacer ajustes en los datos.

Es importante mencionar que cada algoritmo de clasificación tiene sus propios parámetros a ser ajustados. Por lo que, este será un trabajo de gran esfuerzo y paciencia para dar con buenos resultados.

Predicción de la Máquina de Aprendizaje

En esta actividad se utilizará la máquina de aprendizaje automática con nueva información para iniciar la predicción o inferir resultados.

Conclusiones

En el presente trabajo se ha propuesto un proceso denominado MaTexEdu, en él mismo se proponen y adaptan un conjunto de técnicas específicas para la definición de las principales actividades y artefactos, utilizando la ingeniería del dominio en sus disciplinas de análisis, diseño e implementación para ambientes educativos, con fin de obtener una máquina de aprendizaje para el análisis de sentimientos, que permitirá inferir potenciales criterios para ser extrapolados a otros dominios de aplicación.

Es importante destacar que las aplicaciones del procesamiento del lenguaje natural permiten integrar todo el poder comunicativo del habla humana con las capacidades de procesamiento de los sistemas informáticos. Por lo que, la minería de textos es una gran ayuda para las empresas y personas para ser más productivas, a comprender mejor a sus clientes y a utilizar conocimientos para tomar decisiones basadas en datos.

Muchas tareas repetitivas y que consumen mucho tiempo ahora se pueden reemplazar por algoritmos que aprenden de los ejemplos para lograr resultados más rápidos y altamente precisos. Desde donde, las posibilidades de analizar grandes conjuntos de datos y utilizar diferentes técnicas, conduce a observaciones esclarecedoras sobre lo que los clientes piensan y sienten sobre un producto. Así mismo, es interesante establecer posibles mecanismos de integración de la máquina de aprendizaje obtenida, con la plataforma de aplicaciones de la universidad.

Referencias

[1] Hearst, M., Tory, M., and Setlur, V. Toward Interface Defaults for Vague Modifiers in Natural Language Interfaces for Visual Analysis, IEEE Vis2019, Short Papers, Vancouver, 2019.
[2] Software & Systems Process Engineering Metamodel 2.0
[3] R. Canelón. InDoCaS: A process for domain engineering in software production lines with a quality approach. IEEE 2019, CHILECON 2019, Santiago de Chile.
[4] Tomás Baviera, "Técnicas para el Análisis del Sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength". Revista DÍGITOS N° 3, ISSN: 2444-0132 (2017), pp. 33/50.
[5] Abdelrahim Kasem Ahmad, Assef Jafar and Kadan Aljoumaa. Customer análisis de sentimientos prediction in telecom using máquinas de aprendizaje in big data platform. J Big Data (2019) 6:28. https://doi.org/10.1186/s40537-019-0191-6 Springer open
[6] Valence Aware Dictionary and Entiment Reasoner. MIT por C.J. Hutto y Eric Gilbert. 2014.