Ciencias de la Información ISSN: 1606-4925 Vol. 52, No.3 , sept-dic, pp. 3 - 7, 2021

ARTÍCULOS

Modelo de clasificación de información medioambiental para internet basado en un clasificador de texto linear

Environmental information classification model for the internet based on a linear text classifier

Liz Pérez Martínez
Diana Lenia Alfonso Pérez
Manuel Alejandro Naranjo Rey
Eduardo Javier Berrio Turiño
Dianelys Nogueira Rivera

Resumen

Introducción: *en la actualidad, los sitios web de noticias comparten información diariamente. Con el constante desarrollo de las tecnologías de la información, la cantidad de datos que circulan en la red crecen exponencialmente día a día.* ***Objetivo****: construir un modelo de clasificación de información medioambiental mediante la clasificación de textos.* ***Materiales y métodos****: la propuesta está basada en un modelo de clasificación de información a través del clasificador de descenso de gradiente estocástico con el modelo de clasificación linear Máquinas de Soporte Vectorial, se empleó la lematización para reducir los datos que se utilizaron para el entrenamiento. Se utilizó Bag of Words para reducir la dimensión de texto mediante las palabras vacías para el español y crear el vector de características.* ***Resultados y discusión****: se obtuvo un modelo de clasificación de información con un desempeño superior al 90%.* ***Conclusiones****: se validó el modelo creado y su rendimiento con tres diferentes medidas: Exactitud, Medida F1 y la matriz de confusión. Se demostró la importancia de distribuir equitativamente la cantidad de documentos de entrenamiento que pertenecen a cada categoría.*

Palabras clave: *modelo de clasificación; clasificación de textos; máquinas de soporte vectorial; clasificador de noticias*

Abstract

Introduction*: at the present time, news websites share information on a daily basis. With the constant development of information technologies, the amount of data circulating on the network grows exponentially every day.* ***Objective:*** *to build a model for the classification of environmental information by classifying texts.* ***Materials and methods:*** *the proposal is based on an information classification model through the stochastic gradient descent classifier with the Vector Support Machines linear classification model. Lemmatization was used to reduce the data used for the training. Bag of Words was used to reduce the text dimension using the stop words for Spanish and create the feature vector.* ***Results and discussion:*** *an information classification model was obtained with a performance greater than 90%.* ***Conclusions:*** *the created model and its performance were validated following three different measures: Accuracy, Measure F1 and Confusion Matrix. The importance of equitably distributing the amount of training documents that belong to each category was demonstrated.*

Keywords*: classification model; text classification; vector support machines; news classifier*

Introducción

El desarrollo de las tecnologías de la información ha contribuido al incremento exponencial de la cantidad de información disponible en internet. Los sitios webs de noticias se han vuelto una de las principales plataformas de consulta. Sin embargo, si bien existe una gran cantidad de información disponible, es una realidad que no siempre la búsqueda de la misma arroja los resultados deseados. En este sentido, los clasificadores de texto de noticias pueden manejar todo el texto de manera rápida, a la vez que realiza predicciones acertadas de las etiquetas de clasificación que le corresponde a cada noticia.

Existen diversos trabajos sobre la construcción de clasificadores

de noticias. Entre ellos se encuentran trabajos realizados por Zhenzhong Li (2016) y Christian Guardiola Gonzáles (2019), donde se propone un clasificador de noticias basado en asignación de Dirichlet latente y otro empleando Bag of Words, respectivamente. No se aprecia en la bibliografía consultada ningún clasificador enfocado específicamente a clasificar noticias medioambientales. Las categorías seleccionadas para el clasificador han sido basadas en los aspectos de interés reflejados en la Estrategia Ambiental Nacional (2016).

Entre las tecnologías empleadas en esta investigación se destaca Python como lenguaje de programación, dado que ofrece una amplia gama de librerías que sirvió de apoyo al momento de crear nuestro clasificador, se ha seleccionado Scikit-Learn (2020) y Natural Lenguaje Toolkit (NLTK, 2020) puesto que contienen todo tipo de algoritmos para entrenar modelos de clasificadores. De igual manera se emplea Pickle para guardar el modelo obtenido luego del entrenamiento, de modo que pueda ser utilizado posteriormente. Existen alternativas online que nos ofrecen crear un clasificador de texto sin programar. Ejemplo de esto es Microsoft Azure y Amazon Machine Learning (Sitio big data, 2019), el gran inconveniente de estas herramientas es que son de pago y tienen un costo elevado, lo cual va en contradicción con la política del país referida al empleo de software libres.

Existe una amplia variedad de algoritmos de clasificación de textos, entre ellos destacan Máquina de Soporte Vectorial, Naive Bayes, Árboles de Decisión y Regresión Logística (MonkeyLearn, 2020). En esta investigación se empleó Máquinas de Soporte Vectorial (SVM del inglés Support Vector Machines). Entre las razones por las que se seleccionó SVM, se encuentra que la función de decisión utiliza un subconjunto de puntos de entrenamiento llamados vectores de apoyo; por lo tanto, es eficiente en memoria. Además, en la práctica, los modelos SVM son generalizados, con menos riesgo de sobreajuste. De igual manera, funciona muy bien para la clasificación de texto y para encontrar el mejor separador linear.

Para aumentar la efectividad del modelo se emplearon dos métodos matemáticos: Bag of Words y Stochastic Gradient Descent Classifier junto a la lematización. La lematización es el proceso de agrupar las diferentes formas de inflexión de una palabra para que puedan analizarse como un solo elemento. La lematización es similar a la derivación, pero aporta contexto a las palabras (GeeksForGeeks, 2018). Esta técnica evita la aparición de tokens o ítems lingüísticos con un significado similiar pero que son diferentes sintácticamente. Este proceso está basado en los lemas de las palabras, el cual es el representante de todas las formas flexionadas de una misma palabra.

Bag of Words (BoW) hace referencia a una estructura de almacenamiento donde cada artículo es descompuesto y considerado una lista de palabras a modo de diccionario. De esta manera la importancia reside en la frecuencia de las palabras por documento (López Pacheco, 2018). La premisa es que los documentos son similares si tienen un contenido similar.

La clasificación de noticias consiste en analizar el texto de dicha noticia y asignar la categoría a la cual pertenece la misma, con un enfoque de que una noticia puede pertenecer a una sola categoría. Para resolver este problema, se propuso un modelo de clasificador de texto linear. En esta investigación, el modelo utilizó un clasificador linear basado en Descenso de Gradiente Estocástico (SGD, del inglés Stochastic Gradient Descent), de acuerdo con (Scikit-Learn, 2020). SGD es una forma simple y eficiente de entrenar clasificadores lineares como las Máquinas de Soporte Vectorial, empleadas en esta investigación. (Medium, 2020)

El objetivo de esta investigación es construir un modelo de clasificador de noticias basado en el algoritmo de Máquinas de Soporte Vectorial. Dicho modelo podrá ser empleado en un futuro para la construcción de diferentes herramientas basadas en la clasificación automática de noticias medioambientales.

En esta investigación se recolectaron datos de noticias medioambientales publicadas en diferentes portales entre los años 2016 - 2021. Se consideraron las diferentes técnicas de pre procesamiento de texto a aplicar para construir un modelo de clasificación de noticias.

Materiales y métodos

Existen diferentes métodos como son el stemming, la lematización, el análisis morfológico, la desambiguación de sentidos e interpretación de las relaciones semánticas. En esta investigación se empleó la lematización como proceso lingüístico, dada su elevada precisión y su análisis del significado de las palabras, para obtener buenos resultados al emplear el clasificador.

Por lo que se considera una investigación mixta, dadas las características de la información a procesar (cualitativa y cuantitativa); con un alcance exploratorio y explicativo. La investigación se realizó sobre una muestra de 452 documentos provenientes de diferentes sitios webs de noticias medioambientales (en la Tabla 2 se muestra la relación de las categorías con la cantidad de documentos analizados), de una población considerada como toda la información ambiental disponible en internet.

Sin embargo, reducir los datos a procesar con el empleo de la lematización no es suficiente. Los datos en texto no pueden ser directamente procesados por los algoritmos puesto que la mayoría de ellos espera vectores con características numéricas con un tamaño corregido. Para realizar el proceso de convertir el texto de un artículo en un vector de características numéricas se empleó Bag of Words.

BoW realiza tres acciones:

- Tokenizar las palabras y asignarle un identificador numérico para cada token, guiándose por lo espacios y los signos de puntuación como los separadores de los tokens.
- Contar la ocurrencia de los tokens en cada documento
- Normalizar y ponderar con tokens de importancia. decreciente que ocurren en la mayoría de las muestras/ artículos.

Las características y las muestras son definidas de la siguiente manera:

- Cada frecuencia de ocurrencia de un token individual (esté normalizada o no) es considerada una característica.

- El vector de todas las frecuencias de los tokens para un documento dado es considerado una muestra multivariante.

Hay ocasiones dónde el cuerpo del documento es extenso y algunas palabras se van a repetir muchas veces, por lo que van a aportar poca información de importancia acerca del contenido del documento. Para volver a pesar la cantidad de características resultantes y eliminar esas palabras que no aportan información muy útil comúnmente se utiliza *tf-idf. Tf-idf* (del inglés *term frecuency-inverse document frecuency*), frecuencia de término -frecuencia inversa de documento, es una medida numérica que expresa cuan relevante es una palabra en un documento. Su fórmula es la siguiente: "tf-idf(t,d)"="tf(t,d)"×"idf(t)" .[10]

La frecuencia de termino tf(t,d) puede ser determinada de varias formas (Scikit-Learn, 2020):

- El número de veces que el término t ocurre en el documento d: tf(t, d) = f(t,d).
- Frecuencias booleanas: tf(t,d) = 1 si t ocurre en d, y 0 si no
- Frecuencia escalada logarítmicamente: tf(t,d) = 1 + log f(t,d) (y 0 si f(t,d)=0)

La frecuencia inversa de documento se obtiene mediante la siguiente fórmula:

$$\text{idf}(t) = \log \frac{n}{1+\text{df}(t)}.$$

donde n es el número total de documentos y df(t) es el número de documentos que contienen el término t.

Esta investigación empleó la estrategia *Bag of Words* para vectorizar los documentos que serán clasificados. Para la normalización y ponderación de los tokens se empleó *tf-idf* con frecuencia de término igual al número de veces que ocurre el término en un documento.

Resultados y Discusión

Stochastic Gradient Descent Classifier
El clasificador implementa primeramente una rutina de aprendizaje de SGD. El algoritmo itera sobre los datos de entrenamiento y para cada uno de ellos actualiza los parámetros del modelo según la siguiente regla:

$$w \leftarrow w - \eta[\alpha \frac{\partial R(w)}{\partial w} + \frac{\partial L(w^T x_i + b, y_i)}{\partial w}]$$

donde n es el rango de aprendizaje que controla el tamaño del salto en el parámetro espacio, L es la función de pérdida que mide el entrenamiento del modelo, R es el término de regularización que penaliza la complejidad del modelo; α>0 es un parámetro que controla la fuerza de regularización, con los parámetros del modelo $w \in \mathbf{R}^m$ y la intercepción $b \in \mathbf{R}$ y (x_1, y_1) el par de datos de entrenamiento (texto, etiqueta).

El rango de aprendizaje n está dado por la fórmula:

$$\eta^{(t)} = \frac{1}{\alpha(t_0+t)}$$

donde t es el tiempo de salto, t_0 es una heurística propuesta por Léon Bottou de manera que las actualiWzaciones iniciales esperadas sean comparables con el tamaño esperado de los pesos (esto asumiendo que la norma de las muestras de entrenamiento es aproximadamente 1 (Scikit-Learn, 2020) .

Flujo básico del modelo

El modelo de clasificación de texto basado en BoW primeramente carga los datos y los reduce mediante la lematización. Se transforman las noticias a un vector de las frecuencias de los tokens en cada documento. Luego el modelo toma los datos de entrenamiento del vector como el valor de entrada del clasificador SGD para entrenarlo y así obtener un buen rendimiento del mismo.

Los pasos se detallan a continuación:

- Leer los datos mediante la clase *load_files() de Scikit-Learn*.
- Preprocesar los datos a partir de la lematización con la clase *WordNetLemmatizer()* del módulo *stem* de la librería NLTK.
- Preparar los datos derivados en un vector de características mediante el algoritmo BoW con la clase *TfidfVectorizer()* de la librería *Scikit-Learn*, que tiene como parámetros una lista de palabras vacías para el español, para eliminar palabras que no tengan relevancia semánticamente.
- Separar los datos en datos de entrenamiento y datos de prueba con la función t*rain_test_split()* del módulo *sklearn.model_selection.*
- Asignar a *train*_x y *train*_y como las características de cada documento y las etiquetas de cada documento respectivamente.
- Definir como entrada para el algoritmo SGD *train*_x y *train*_y para el comienzo del entrenamiento y la obtención del clasificador mediante la clase *SGDClassifier()* del módulo *sklearn,linear_model* a partir del empleo de la función *fit()* para entrenarlo.
- Validar la exactitud del modelo con el empleo de los datos de prueba que fueron separados anteriormente para predecir las categorías de cada documento mediante la función *predict()* del módulo *sklearn,linear_model*.
- Guardar el modelo con la función *dump()* de la librería *Pickle.*

Para validar el modelo de clasificador de notiWcias, se utilizó la matriz de confusión, la medida F1 y la exactitud. La matriz de confusión permite visualizar el desempeño del algoritmo; cada columna representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. La medida F1 es la media armónica entre la precisión y la exhaustividad. La exactitud mide el porcentaje de casos que el modelo ha acertado (IArtificial, 2020).

Tabla 1. Matriz de confusión
Fuente: elaboración propia

Realidad/Predicción	0	1
0	TN	FP
1	FN	TP

Estas medidas se calculan de la siguiente forma:

- Matriz de confusión (Tabla 1)

Donde:

- TN: True Negative (Negativo Verdadero)
- TP: True Positive (Positivo Verdadero)
- FP: False Positive (Positivo Falso)
- FN: False Negative (Negativo Falso)

Positivo o Negativo se refiere a la predicción y Verdadero o Falso se refiere a si la predicción es correcta o no.

- Medida F1

$$F1 = \frac{2 * exhaustividad * precisión}{exhaustividad + precisión}$$

- Exactitud

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

Los datos a partir de los que fue construido el modelo provienen de diferentes sitios webs de noticias medioambientales. Estos datos se pueden clasificar en seis categorías diferentes: agua, bosques, calidad ambiental, cambio climático, diversidad biológica y suelos. En la Tabla 2 se muestra la relación de las categorías con la cantidad de documentos. Estos datos se dividieron en datos de entrenamiento y datos de prueba, tomándose para los últimos el veinte por ciento del total de documentos.

Tabla 2. Datos para construir el modelo
Fuente: elaboración propia

Identificador Clase	Nombre de la Clase	Número de documentos
0	Agua	70
1	Bosques	68
2	Calidad ambiental	78
3	Cambio Climático	94
4	Diversidad Biológica	75
5	Suelos	67

Podemos observar en la Figura 1 la que el modelo alcanzó el 91% de exactitud. La exactitud representa el número de predicciones correctas realizadas por el modelo. Se considera que la mejor precisión es el 100% y que la cantidad de artículos por categoría está representada equitativamente podemos afirmar que el resultado de la exactitud refleja un buen rendimiento por parte del modelo.

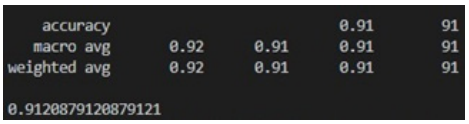

Figura 1. *Porcentaje de exactitud del modelo.*
Fuente: *elaboración propia*

En la Tabla 3 se observan los valores que tomó la medida F1. Como se ha mencionado anteriormente, la misma es el promedio ponderado entre la precisión y la exhaustividad, por lo que considera tanto los falsos positivos como los falsos negativos. De igual manera, al estar distribuidos equitativamente los artículos de entrenamiento por cada categoría del clasificador, se afirma que los resultados de la Medida F1 para cada categoría reflejan un buen rendimiento por parte del modelo.

Tabla 3. Valores de la Medida F1
Fuente: elaboración propia

Categoría	Precisión	Exhaustividad	Medida F1
0	0.86	0.92	0.89
1	1	0.85	0.92
2	1	0.98	0.94
3	0.89	1	0.94
4	0.79	0.94	0.86
5	1	0.87	0.93

En la Tabla 4 se observan los valores que tomó la matriz de confusión. Al analizar los valores obtenidos se aprecia que solo erró en la clasificación de ocho artículos de un total de 91 que se utilizaron para hacer las pruebas. Por lo tanto, al haber acertado aproximadamente el 91% del total de artículos, demuestra que el modelo presenta un buen desempeño.

Tabla 4. Valores de la Matriz de Confusión
Fuente: elaboración propia

	0	1	2	3	4	5
0	12	0	0	0	1	0
1	0	11	0	1	1	0
2	1	0	16	1	0	0
3	0	0	0	16	0	0
4	1	0	0	0	15	0
5	0	0	0	0	2	13

Conclusiones

Con la aplicación del modelo de clasificador de noticias medioambientales, se contribuye a facilitar la búsqueda de la información referente a esta temática disponible en internet. El estudio de los clasificadores existentes arrojó que estos no responden a informaciones de carácter medioambiental. El modelo del clasificador permite tener toda la información estructurada, de manera que se viabilice el acceso a la misma. Las pruebas

realizadas al modelo mediante diferentes medidas permiten apreciar la relevancia de distribuir los datos de entrenamiento equitativamente entre cada categoría del clasificador. La validación del modelo demostró el eficiente desempeño del mismo, con una exactitud en la clasificación por encima del 90%. Este modelo que deberá reentrenarse cada cierto período de tiempo puesto que al comenzar a utilizarlo el rendimiento comenzará a decrecer, este periodo estará determinado por su uso.

Referencias

Estrategia Ambiental Nacional (2016). CITMA. http:// http://repositorio.geotech.cu/jspui/bitstream/1234/2727/1/Estrategia%20Ambiental%20Nacional%202016-2020.pdf

GeeksForGeeks. (2018, November). Python | Lemmatization with NLTK. https://www.geeksforgeeks.org/python-lemmatization-with-nltk/

Guardiola González, C. (2019). Clasificador de texto mediante técnicas de aprendizaje automático [Trabajo de Fin de Grado, Universidad Politécnica de Valencia] https://riunet.upv.es/bitstream/handle/10251/133840/Guardiola%20-%20Clasificador%20de%20textos%20mediante%20t%C3%A9cnicas%20de%20aprendizaje%20autom%C3%A1tico.pdf?sequence=1&isAllowed=y

IArtificial. (2020, octubre). Precision, Recall, F1, Accuracy en clasificación. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

López Pacheco, Iván. (2018) Análisis comparativo de algoritmos de Deep Learning para la clasificación de textos [Trabajo de Fin de Grado, Universidad Carlos III de Madrid] https://e-archivo.uc3m.es/bitstream/handle/10016/29209/TFG_Ivan_Lopez_Pacheco_2018.pdf?sequence=1

Medium (2020, septiembre). Máquina de Soporte Vectorial (SVM). https://medium.com/@csarchiquerodriguez/maquina-de-soporte-vectorial-svm-92e9f1b1b1ac

MonkeyLearn. (2020, agosto) 5 Types of Classification Algorithms in Machine Learning https://monkeylearn.com/blog/classification-algorithms/

NLTK.(2020) Natural Language Toolkit https://www.nltk.org

Scikit-Learn. (2020) Feature extraction https://scikit-learn.org/stable/modules/feature_extraction.html

Scikit-Learn. (2020) https://www.scikit-learn.org/stable

Scikit-Learn. (2020) Stochastic Gradient Descent https://scikit-learn.org/stable/modules/sgd.html#:~:text=Stochastic%20Gradient%20Descent%20(SGD)%20is,Vector%20Machines%20and%20Logistic%20Regression.&text=The%20advantages%20of%20Stochastic%20Gradient,Efficiency.

Sitio big data. (2019). Amazon Machine Learning, Azure, Cloud AI y Watson. https://sitiobigdata.com/2019/01/31/mlaas-amazon-machine-learning/#

Zhenzhong, Li. (2016). News text classification model based on topic model https://www.researchgate.net/publication/306925662_News_text_classification_model_based _on_topic_model. 10.1109/ICIS.2016.7550929

Recibido: 18 de octubre de 2021
Aprobado en su forma definitiva:
20 de diciembre de 2021

Liz Pérez Martínez
Universidad de Matanzas, Matanzas, Cuba.
Correo-e.: lizy.perez@umcc.cu

Diana Lenia Alfonso Pérez
Universidad de Matanzas, Matanzas, Cuba.
Correo-e.: diana.alfonso@est.umcc.cu

Manuel Alejandro Naranjo Rey
Universidad de Matanzas, Matanzas, Cuba.
Correo-e.: manuel.naranjo@umcc.cu

Eduardo Javier Berrio Turiño
Universidad de Matanzas, Matanzas, Cuba.
Correo-e.: eduardo.berrio@umcc.cu

Dianelys Nogueira Rivera
Universidad de Matanzas, Matanzas, Cuba.
Correo-e.: dianelys.nogueira@umcc.cu