ISSN 1870-4069

Clasificación automática de textos estructurados utilizando inteligencia artificial

Michelle Alejandra Meléndez-Cardiel, Rafael Guzmán-Cabrera

Universidad de Guanajuato,
México

{ma.melendezcardiel, guzmanc}@ugto.mx

Resumen—Una de las tareas más comunes del Data Mining es la clasificación de textos de acuerdo con sus características, para ello es necesario el uso de un corpus con el cual se pueda trabajar para el diseño de un modelo de clasificación dependiendo de laspalabras que contenga cada clase. Con ese fin, se necesita saber discriminar qué etiquetas o cualidades nos serán de utilidad para dicha categorización. Para este trabajo se presenta un análisis con 3 diferentes clasificadores más usados: Support Vector Machine (SVM), Naive Bayes y Regresión Logística. Los experimentos se prueban haciendo uso de 6 carpetas de las 20 que contiene The 20 Newsgroups como corpus, con su tema dirigido a la política y religión; con un 10 por ciento de pruebas. Haciendo uso de Area bajo la curva, Precisión, Recall y Macro-F1 en este trabajo, ya que esas métricas de evaluación son mejores para demostrar el rendimiento del modelo.

Palabras clave: clasificación de textos, inteligencia artificial, minería de datos, decisión.

Automatic Classification of Structured Texts Using Artificial Intelligence

Abstract. One of the most common tasks of Data Mining is the classification of texts according to their characteristics, for this it is necessary to use a corpus with which one can work to design a classification model depending on the words that are used. contain each class. To this end, it is necessary to know how to discriminate which labels or qualities will be useful for said categorization. For this work, an analysis is presented with 3 different most used classifiers: Support Vector Machine (SVM), Naive Bayes and Logistic Regression. The experiments are tested using 6 folders of the 20 that the 20 Newsgroups contains as corpus, with its topic directed to politics and religion; with 10 percent testing. Making use of Area under the curve, Precision, Recall and Macro-F1 in this work, since those evaluation metrics are better to demonstrate the performance of the model.

Keywords: Text classification, artificial intelligence, data mining, decision.

1. Introducción

Las técnicas de inteligencia artificial están destinadas a ayudar a tomar decisiones en escenarios donde la cantidad deinformación complica el análisis desarrollado por los expertos. El análisis del comportamiento de los consumidores es un problema fundamental para la formulación de estrategias de marketing, ventas, atención al cliente, fidelización, segmentación, cross-selling, optimización de la cadena de suministro, detección y prevención de fraudes y Detecciónde Anomalías (DA), entre otras [1].

La minería de datos se define como la práctica de examinar una gran base de datos preexistente para generar nueva información. Una tecnología robusta tiene un gran potencial para ayudar a las organizaciones a concentrarse en la información más importante en los almacenes de datos almacenados.

Las herramientas y técnicas de minería de datospredecirán las tendencias futuras al hacer que el negocio sea más proactivo y mejores decisiones basadas en el conocimiento. Las técnicas de minería de datos podrían responder preguntas relacionadas con el negocio que tradicionalmente requerían demasiado tiempo para resolver [2].

La clasificación de texto permite el agrupamiento de documentos semánticamente significativos, ayudando tanto a losusuarios como a las herramientas de recuperación de información, a localizarlos con mayor precisión. En esta sección se compara el rendimiento en términos de precisión de la clasificación [3].

2. Antecedentes

The 20 Newsgroups Dataset se divide en 20 carpetas, cada una con mil archivos aproximadamente, sobre 20 temas. Consta de 20000 mensajes tomados de 20 grupos de noticias. Los mensajes son proporcionados por la Escuela de Informática de la Universidad Carnegie Mellon, que se remonta a 1999.

Los 20000 mensajes representan 1000 artículos de uso neto con aproximadamente el 4% de los artículos publicados de forma cruzada, son publicaciones típicas, por lo que tienen encabezados con líneas de asunto, firma archivos y partes citadas de otros artículos, tambiéncada grupo de noticias se almacena en un subdirectorio, con cada artículo almacenado como un archivo separado.

Este conjunto de datos viene organizado por fecha, y también viene con algo de ruido (como los encabezados "De", "Asunto" en cada publicación como se mencionó anteriormente y también algunos errores tipográficos en el conjuntode datos); algunos de los temas están muy relacionados entre sí, como Hardware MAC y Hardware PC, mientras queotros temas son muy diferentes, como temas cristianos y motocicletas [4].

Este corpus ha sido muy utilizado para diversos trabajos anteriormente, algunos usan todas las carpetas con los mil archivos de cada una, a diferencia de otros únicamente utilizan 6 carpetas, con 2 temas diferentes relacionados y solo100 archivos de cada carpeta. Para identificar de mejor manera cada trabajo, se presenta a continuación en la Tabla1 la comparación de cada trabajo.

Tabla 1. Comparativa de resultados obtenidos en algunos trabajos recabados.

	X2	MI	TF	Precisión	Recall	F1	
MNB	93.16%	93.29%	93.36%	-	-	-	[5]
	-	-	-	86%	-	-	[6]
	-	-	-	72%	70%	70%	[7]
	79.15%	79.07%	-	-	-	79.29%	[6]
SVM	93.16%	93.29%	93.36%	-	-	-	[7]
				71%	67%	67%	[7]

2.1 Técnicas de clasificación

Naïve Bayes es un método de clasificación supervisado y generativo que se basa en el teorema de Bayes y en lapremisa de independencia de los atributos1 para obtener la probabilidad de que un documento pertenezca a una determinada clase como se indica en la ecuación que sigue:

$$P(C_i|D) \text{ a } P(C_i) \text{ G } \prod_{k=1}^{n} P(f_k|C_i), \quad (1)$$

donde f_k son los atributos del documento, C_i es la clase y $P(f_k|C_i)$ es la probabilidad de ocurrencia del atributo enla clase dada. La clase seleccionada por el clasificador será la que maximice la probabilidad anterior. Las implementaciones del algoritmo de Naïve Bayes difieren principalmente en la aproximación de $P(f_k|C_i)$ y las técnicas de smoothing utilizadas para el tratamiento de probabilidades bajas o nulas [8].

Support Vector Machines es un método supervisado de clasificación binaria en el cual el entrenamiento consiste en encontrar un hiperplano que separe los vectores de atributos que representan los documentos del conjunto de datos en dos grupos, siendo esta separación la más grande posible. Aquellos vectores que definen los márgenes de lamáxima separación entre las clases se conocen como support vectors.

Para la predicción de la clase utilizando este modelo se define la ecuación que sigue:

$$f(x) = sign\left(\sum_i \alpha_i x_i \cdot x + b\right). \quad (2)$$

Siendo, x el vector de atributos del documento a clasificar, α_i cada uno de los pesos que ponderan los vectores de atributos identificados como support features, x_i cada uno de los support features y b el termino independiente. Unvalor de —1 indicara que el documento pertenece a una clase y un valor de +1 a la otra, lo que representa de qué lado del hiperplano se encuentra x [8].

La **Regresión Logística** (también conocido como clasificador de máxima entropía), es un modelo matemático utilizado para predecir el resultado de una variable categórica, por lo general dicotómica, en función de las variables independientes o predictoras. La predicción que se obtiene es la probabilidad de pertenecer a cada clase. Una de las ventajas fundamentales de la regresión logística sobre otras técnicas, es que

el resultado del modelo entrenado se puede interpretar fácilmente. Esto se debe a que el coeficiente obtenido para cada variable dependiente, indica de qué manera influye en el modelo dicha variable. Otras ventajas son su simplicidad y eficacia [9].

2.2 Métricas de evaluación

The components update was according to functionality and ease of configuration, they are listed below, see figure 1.

Dos de las medidas para medir el rendimiento son la precisión y el recall. La **precisión** de un clasificador se definecomo la fracción de etiquetas correctamente asignadas entre todas las etiquetas asignadas por clasificador:

$$Precisión = \frac{TP}{TP+TP}.$$

El **recall** de un clasificador es la fracción de etiquetas correctamente clasificadas entre todas las etiquetas realmente positivas:

$$Recall = \frac{tp}{tp+fn},$$

t_p indica el número de instancias positivas clasificados correctamente,

f_p las instancias incorrectamente asignadas o falsos positivos,

f_n las instancias no asignadas o falsos negativos [10].

Área bajo la curva (AUC por sus siglas en inglés) se calcula usando el área bajo la curva ROC y cuanto mayor esel área más precisa es formalmente el predictor, la fórmula para calcular el AUC es representada por:

$$AUC = \int_0^1 f(x)dx,$$

donde f(x) representa la función de la curva característica de funcionamiento del receptor (ROC por sus siglas en ingles), sin embargo, desde f(x) tiende a no tener una forma de integración como una parábola; varios autoressugieren utilizar métodos de aproximación para calcular las AUC.

F1 es una medida de precisión en una prueba que se calcula a partir de la precisión y el recall de la prueba que seestá llevando a cabo, en pocas palabras F1 es la media armónica de la precisión y el recall, que se muestra a continuación: [11]:

$$F1 = \frac{t_p}{t_p + \frac{1}{2}(f_p+f_n)}.$$

3. Metodología

Para realizar este trabajo se desarrolló la clasificación de documentos pertenecientes a The 20 Newsgroups Dataset, los cuales tienen 20 temas distintos de noticias, un tópico por carpeta, desde deportes y motocicletas, hasta política yreligión. Debido a que 3 carpetas son pertenecientes al tema de política: “talk.politics.guns”, “talk.politics.misc” y “talk.politics.mideast” y 3 carpetas corresponden a religión: “alt.atheism”,

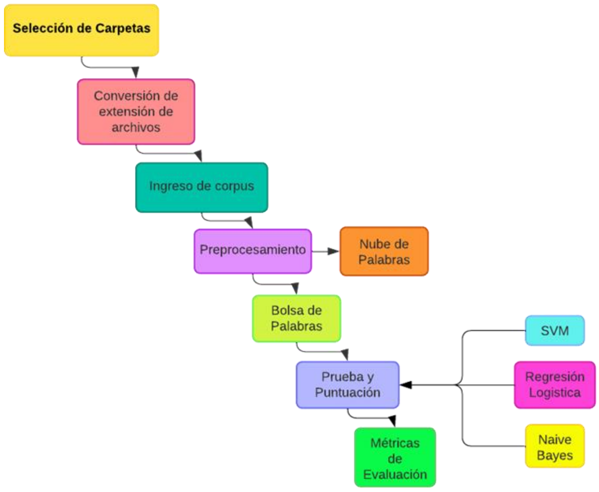

Fig. 1. Metodología de los experimentos realizados Se genera una bolsa de palabras con los archivos procesados y se realiza un análisis para discriminar los que tienen resultados bajos y elegir el que tenga mayor éxito de precisión, para esto se utilizaron los métodos de aprendizaje: Support Vector Machine (SVM), Regresión Logistica (RL) y Naive Bayes (NB), apoyados por las métricas de evaluación: Area bajo la curva (AUC), Recall, F1 y precisión.

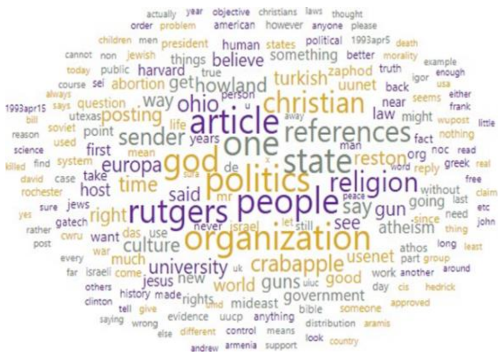

Fig. 2. Nube de palabras del proceso.

“soc.religion.christian”, “talk.religion.misc”, se optó por trabajar con estas seis carpetas y lograr minimizar el trabajo del equipo con el que setrabajó. En la Figura 1, se muestra el diagrama que señala la metodología implementada para realizar este estudio.

A continuación, se describen los componentes pertenecientes a la metodología. Para el ingreso de datos se tuvo que convertir su extensión a txt, ya que se desconocía el tipo de archivo que era. Posteriormente realizando los experimentos en Python, se le aplicó un preprocesamiento eliminando las stopwords, signos de puntuación y convirtiendo a minúsculas las palabras.

Finalizando esto, se crea una Nube de Palabras como se muestra en la Figura 2, para apoyar la mayor cantidad de menciones de palabras principales de los temas. Se genera una bolsa de palabras con los archivos procesados y se realiza un análisis para discriminar los que tienen resultados bajos y elegir el que tenga mayor éxito de

Tabla 2. Resultados del Test and Score de la carpeta con archivos crudos.

Modelo	AUC	F1	Precisión	Recall
SVM	0.805	0.418	0.766	0.444
Naive Bayes	0.956	0.805	0.819	0.807
Regresión Logística	0.965	0.871	0.871	0.871

Tabla 3. Resultados del Test and Score de la carpeta con archivos crudos sin stopwords.

Modelo	AUC	F1	Precisión	Recall
SVM	0.854	0.377	0.744	0.416
Naive Bayes	0.973	0.877	0.878	0.877
Regresión Logística	0.973	0.871	0.871	0.871

Tabla 4. Resultados del Test and Score de la carpeta sin etiquetas.

Modelo	AUC	F1	Precisión	Recall
SVM	0.602	0.130	0.494	0.206
Naive Bayes	0.908	0.634	0.680	0.629
Regresión Logística	0.922	0.734	0.736	0.734

Tabla 5. Resultados del Test and Score de la carpeta sin etiquetas sin stopwords.

Modelo	AUC	F1	Precisión	Recall
SVM	0.673	0.224	0.783	0.262
Naive Bayes	0.966	0.729	0.830	0.717
Regresión Logística	0.970	0.864	0.865	0.864

precisión, para esto se utilizaron los métodos de aprendizaje: Support Vector Machine (SVM), Regresión Logistica (RL) y Naive Bayes (NB), apoyados por las métricas de evaluación: Area bajo la curva (AUC), Recall, F1 y precisión.

4. Resultados

De donde se descargó la base de datos de The 20 Newsgroups, contenía 2 carpetas, la primera eran los archivos completamente crudos sin modificar nada y la segunda donde se eliminaban las etiquetas que no se utilizaban de todoslos archivos. Al observar esto, se optó por hacer el mismo experimento con ambas carpetas, es decir, probar los resultados cuando los datos se ingresan sin modificar y cuando se eliminan las stopwords.

Como cada carpeta tiene mil archivos aproximadamente, se realizaron los experimentos con los seis mil archivos,pero debido a que eran muchos archivos para el equipo con el que se trabaja, se optó por únicamente tomar 250 archivos de cada carpeta, teniendo un total de 1500 archivos.

De los experimentos se obtuvieron los resultados que se muestran en los cuadros a continuación, donde se puede observar que la carpeta con mejores resultados es la que contiene los archivos crudos y se le aplica la eliminación de stopwords (Tabla 3), la cual es con la que se trabajará para los siguientes experimentos.

5. Conclusión

Para la elaboración de este trabajo, se observa que se obtienen mejores resultados cuando se hace uso de las carpetas con todas las etiquetas de información de cada archivo, pero eliminando las stopwords, logrando alcanzar el 97.3% con la métrica de evaluación de área bajo la curva y los métodos de aprendizaje Naïve Bayes y Regresión Logística.

Haciendo la comparación de los resultados obtenidos en las tres investigaciones que se recabaron y que se muestranen el presente trabajo, se logró superar los resultados, teniendo en cuenta que es el mismo dataset y la misma metodología, variando el preprocesamiento a pesar de utilizar únicamente 3 métodos de aprendizaje y 4 métricas de evaluación, siendo posible trabajar con otros existentes.

Referencias

1. Escobar, H., Alcivar, M., Puris, A.: Aplicaciones de minería de datos en marketing. Revista Publicando, vol. 3, no. 8, pp. 503–512 (2016)
2. Sidow-Osman, A.: Data Mining Techniques: Review. International Journal of Data Science Research, vol. 2, no. 1, pp. 1–4 (2019)
3. Sánchez-Vera, M. D. M.: El pensamiento computacional en contextos educativos: una aproximación desde la Tecnología Educativa. Research in Education and Learning Innovation Archives, pp. 24–29 (2019) doi: 10.7203/realia.23.15635
4. Akef, I., Arango, J. S. M., Xu, X.: Mallet vs GenSim: Topic modeling for 20 news groups report. University of Arkansas at Little Rock, pp. 1–12 (2016)
5. Chandra, A.: Comparison of feature selection for imbalance text datasets. In: 2019 International Conference on Information Management and Technology (ICIMTech) IEEE., vol. 1, pp. 68–72 (2019) doi: 10.1109/ICIMTech.2019.8843773
6. Adi, A. O., Çelebi, E.: Classification of 20 news group with Naïve Bayes classifier. In: 2014 22nd Signal Processing and Communications Applications Conference (SIU), IEEE, pp. 2150–2153 (2014) doi: 10.1109/SIU.2014.6830688
7. Dubiau, L., Ale, J. M.: Análisis de sentimientos sobre un corpus en español: experimentación con un caso de estudio. In: XIV Argentine Symposium on Artificial Intelligence (ASAI) – JAIIO, vol. 42, pp. 36–47 (2013)
8. Rosenbrock, G., Trossero, S., Pascal, A.: Técnicas de análisis de sentimientos aplicadas a la valoración de opiniones en el lenguaje español. In: XXVII Congreso Argentino de Ciencias de la Computación (CACIC), pp. 291–300 (2021)
9. Poccohuanca, Q., Edmit, O.: Integración de técnicas de deep learning y algoritmos de aprendizaje multi etiqueta para la Clasificación de Textos (2018)
10. Morales-Castro, J. C., Ledesma-Carrillo, L. M., Guzman-Cabrera, R.: Identificación de polaridad en Twitter usando validación cruzada. Identidad Energetica, vol. 4, pp. 87-91 (2022) https://www.researchgate.net/publication/357860394_Identificacion_de_polaridad_en_Twitter_usando_validacion_cruzada