Revista de I+D TECNOLÓGICO
Portal de revistas: http://revistas.utp.ac.pa

Aprendizaje automático aplicado al análisis de sentimientos

Machine learning applied to the sentiment analysis

Denis Cedeno-Moreno [1], Miguel Vargas-Lombardo[2]*
[1] GISES CIDITIC, Facultad de Ingeniería de Sistemas Computacionales, Universidad Tecnológica de Panamá, Panamá
***Autor de correspondencia:** denis.cedeno@utp.ac.pa

RESUMEN– Con la evolución del Internet, hay una gran cantidad de información presente en la web como lo son las opiniones de los usuarios o consumidores sobre diversos contextos ya sea para expresar su conformidad o inconformidad sobre un producto o servicio recibido, así como la opinión de un artículo comprado o sobre la gestión que realiza alguna persona. Debido a la gran cantidad de opiniones, comentarios y sugerencias de los usuarios, es muy importante explorar, analizar y organizar sus puntos de vista para tomar mejores decisiones. El análisis de sentimientos es una tarea de procesamiento de lenguaje natural y extracción de información que identifica las opiniones de los usuarios explicadas en forma de comentarios positivos, negativos o neutrales. Varias técnicas pueden ser utilizadas para este fin, por ejemplo, el uso de diccionarios léxicos que ha sido muy utilizada y recientemente la utilización de la inteligencia artificial específicamente algoritmos supervisados. En este documento, se propone la utilización de técnicas de algoritmos supervisados para observar su utilización y ver el rendimiento de diferentes modelos de algoritmos supervisados para medir la efectividad en la clasificación de un conjunto de datos.

Palabras clave– *Análisis de sentimientos, inteligencia artificial, aprendizaje automático, procesamiento de lenguaje natural.*

ABSTRACT– With the evolution of the Internet, there is a large amount of information present on the web such as the opinions of users or consumers about different contexts, either to express their agreement or disagreement about a product or service received, as well as the opinion of a item purchased or about the management performed by someone. Due to the large number of opinions, comments and suggestions from users, it is very important to explore, analyze and organize their views to make better decisions. Sentiment analysis is a natural language processing and information extraction task that identifies the opinions of the users explained in the form of positive, negative or neutral comments. Several techniques can be used for this purpose, for example the use of lexical dictionaries that has been widely used and recently the use of artificial intelligence specifically supervised algorithms. In this document, we propose the use of supervised algorithm techniques to observe their use and see the performance of different models of supervised algorithms to measure the effectiveness in the classification of a data set.

Keywords– *Artificial intelligence, machine learning, natural language processing, Sentiment analysis.*

1. Introducción

En gran medida por la explosión en Internet de la Web 2.0, nos hemos encontrado con un crecimiento abrumador de la información disponible. Así mismo el acelerado y rápido crecimiento del contenido generado por los usuario por ejemplo redes sociales, blogs o foros, en donde encontramos que está cargado de diferentes opiniones de los usuarios sobre diferentes aspectos cotidianos y dichas opiniones también conocidas como información subjetiva logran generar un sentimiento en las personas [1].

En un texto la información subjetiva tiene un gran potencial. La información subjetiva contiene el punto de vista de la persona que la expone en la cual influye sus intereses y deseos. Esta información puede ser procesada por organizaciones públicas o privadas para conocer más sus estrategias y proyectarse mejor.

La lingüística computacional es una de las áreas que se ha enfocado en el estudio de las opiniones, o mejor dicho del lenguaje subjetivo. Este tipo de lenguaje se emplea para expresar estados personales en el contexto de una conversación o un texto [2].

Por ejemplo, la información subjetiva se puede emplear para, conocer la opinión de los clientes a cerca de un producto o servicio o saber la opinión que tiene el electorado sobre un posible candidato de elecciones, conocer que producto prefieren los clientes [3].

Por estas razones expuestas anteriormente hoy día existe un gran interés por parte de los investigadores y desarrolladores de software por la minería de opiniones

Citación: D. Cedeño y M. Vargas, "Aprendizaje automático aplicado al análisis de sentimientos", *Revista de I+D Tecnológico*, vol. 16, no. 2, pp. (no_modificar), 2020.
Tipo de artículo: Original. **Recibido:** 2 de septiembre de 2019. **Recibido con correcciones:** 12 de septiembre de 2019. **Aceptado:** 8 de julio de 2020.
DOI.

(MO), también llamada análisis de sentimientos (AS) [4], la cual es una tarea del procesamiento del lenguaje natural (PLN) que permite identificar las opiniones relacionadas con un objeto dentro de un contexto común [5].

Para [6] las opiniones son una parte importante en las decisiones del ser humano, cuando una persona desea tomar una decisión se basa en los comentarios de otras personas, por ejemplo, para comprar un producto, seleccionar un destino turístico, incluso para votar por un partido político.

Comúnmente el AS se aplica extrayendo revisiones para un dominio específico, así por ejemplo puede ser de un producto, película, música, un candidato. El AS se puede realizar en diferentes niveles del texto que a lo que se conoce como granularidad, puede ser del documento completo, por oración o por características. En cuanto al documento o por oración el AS tiene como finalidad clasificar la orientación general del sentimiento que se expresa [7].

El concepto de AS no es nuevo, desde hace varios años, los especialistas han ido preparando cantidades innumerables de conjuntos de datos etiquetados con polaridades que tienen etiquetas con simples positivos y negativos hasta conjunto de datos mucho más elaborados que determinan el grado de positividad o negatividad que tiene un texto estudiado.

El AS [8] se apalanca en la utilización de una gran variedad de herramientas tecnológicas para lograr su objetivo, el cual es poder determinar el tono emocional que tiene una palabras.

La mayoría de los proyectos desarrollados para AS han basado su estrategia en la categorización manual. Otros pocos han automatizado el análisis del texto y utilizan diccionarios léxicos para etiquetar la palabra [9].

La inteligencia artificial (IA) que permite crear sistemas expertos, ha evolucionado el concepto del aprendizaje automático. Con la IA se están realizando proyectos de análisis de opiniones trabajando con aprendizaje automático, sin embargo, este tipo de soluciones son muy escasas.

El objetivo principal de este trabajo consiste en aplicar modelos del aprendizaje automático en una solución de AS, sobre un conjunto de datos o corpus, que se ha obtenido del Taller de análisis de sentimientos en español para ver el mejor rendimiento de estos modelos que permitan ver la polaridad ya sea positiva o negativa de las opiniones emitidas.

El resto de este documento está estructurado de la siguiente manera: Sección 2 presenta os antecedentes. Sección 3 los materiales y métodos. Sección 4 los resultados. Sección 5 se describe la discusión y finalmente en la sección 6 las conclusiones y trabajo futuro.

2. Antecedentes

Cada día una enorme cantidad de datos son generados en las organizaciones, por ejemplo, de salud, por tal motivo es necesario el diseño y desarrollo de nuevas y potentes herramientas de procesamiento de la información, con el avance de las tecnologías relacionadas con la información se pueden acceder y analizar estos datos todos ellos relacionados con los registros electrónicos de salud del paciente.

2.1 Análisis de sentimiento.

La forma de comunicarnos a diario con las demás personas es a través del lenguaje natural (LN). Los lenguajes naturales tienen un gran poder expresivo y su función y valor como una herramienta para razonamiento.

El LN [10] ha venido perfeccionándose a partir de la experiencia a tal punto que puede ser utilizado para analizar situaciones altamente complejas. Todo lo que expresamos en Internet en forma de LN como las opiniones puede ser visto como una forma de información no estructurada.

En la actualidad según [11] se afirma que la información se encuentra clasificada de dos formas: estructurada y no estructurada. Desde el punto de vista del análisis cuantitativo de datos, el LN que usamos los humanos para comunicarnos a menudo es incluido dentro de la categoría de datos no estructurados [12]. Por ello, hoy día dentro de la informática ha logrado mucho auge las tareas que incluyen el procesamiento de estos datos no estructurados [13].

Se estima que entre el 80% y el 90% de la información de las organizaciones es no estructurada. La relación entre estas dos formas de clasificación de la información estructura y no estructurada es clave, debido a que hoy día una elevada cantidad de información está en forma no estructurada y requiere el uso de técnicas de procesamiento automático para poder tratarlos.

Existe gran cantidad de información no estructurada en las organizaciones y esta requiere el uso de técnicas de PLN para poder procesarla [14]. Según recientes investigaciones se estima que, en un minuto, habrá:

- 204 millones de emails
- 2,46 millones de post en Facebook
- 320.000 tweets
- 54.000 post en Tumblr
- 17 artículos nuevos en Wikipedia

El tratamiento automático del lenguaje natural o lenguaje humano se llama PLN. El PLN es la parte de la IA que se define como "la ciencia y la ingeniería de hacer máquinas inteligentes" [15].

El AS, también llamado minería de opiniones o análisis de opiniones, en años recientes ha tomado gran interés e importancia gracias al inmenso auge de las redes sociales en las que compartimos a diario nuestras experiencias son una fuente inagotable opiniones.

Los sistemas tradicionales de aplicación del PLN fueron para el tratamiento de textos que describían hechos o que se pueden observar y comprobar en la realidad. En la actualidad, sin embargo, la información sobre hechos ya no es la principal fuente para extraer conocimiento, debido a que la mayoría de los textos en la Web contiene expresiones con sentimientos y repletos de subjetividad.

Así han surgido en la actualidad nuevos campos de investigación dentro del PLN cuyo propósito es analizar la subjetividad en las opiniones o extraer y clasificar los sentimientos de las opiniones expresadas por las personas en el ámbito educacional, político, social y económico [16].

Dentro del PLN el área encargada de la detección automática de los sentimientos expresados en los textos y su clasificación según la polaridad que tienen (positiva, negativa o neutra) es el área de AS. El AS se puede utilizar para medir la intensidad de una opinión así como también puede ponderar cuantitativamente expresiones que muchas veces nos resultan subjetivas, permitiendo saber si se está hablando de forma positiva, negativa o neutra sobre un amplio contexto como servicios, productos, películas entre otros [17].

Con la importancia actual de las redes sociales, donde a diario se comenta de todo un poco, la posibilidad de tener herramientas que permitan monitorizar y valorar las opiniones vertidas en estas redes sociales son un punto de partida para el análisis de información no estructurada. Las soluciones tecnológicas de AS hacen posible extraer un valor tangible y directo a los comentarios de usuarios emitidos en lenguaje natural.

El AS, se refiere a la aplicación de una serie de técnicas del PLN, lingüística computacional y minería de textos, cuyo objetivo es la extracción de información subjetiva a partir de contenidos generados por los usuarios, como comentarios en blogs u opiniones en revistas de productos [18].

Existen básicamente dos enfoques para el AS, el enfoque basado en diccionarios semánticos y el basado en aprendizaje automático.

El enfoque de diccionarios semánticos hace uso de lexicones, en este enfoque principalmente se etiquetan palabras dando un valor negativo o positivo a cada palabra. Entre los lexicones más utilizados en la literatura se encuentran SentiWordNet y WordNet-affect [19].

El enfoque basado en aprendizaje automático requieren de un conjunto de características para posteriormente entrenar un algoritmo de clasificación [17]. En la figura 1 se muestra el proceso general del aprendizaje automático seguido en este experimento.

Como primer paso tenemos la extracción de características que se hace desde un conjunto de datos de entrenamiento, estas se representan como un vector de características. Estas características son extraídas a través de diversos métodos como bolsa de palabras, n-gramas entre otros. Luego el segundo paso es el entrenamiento del algoritmo de clasificación seleccionado con los vectores de características [20].

En este enfoque el método para la extracción de características y el algoritmo de clasificación seleccionado juegan un rol importante en la precisión de la clasificación de la opinión. Ya que elegir el del algoritmo de aprendizaje y la determinación de las características correctas es fundamental para obtener un buen resultado en la clasificación.

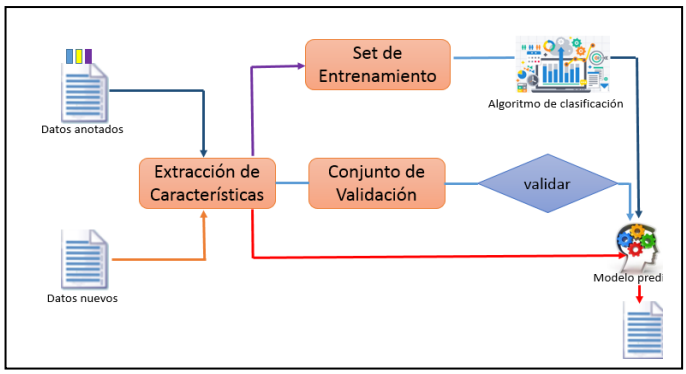

Figura 1. Proceso de aprendizaje automático.

2.2 Aprendizaje Automático.

El aprendizaje automático o machine learning [21], es una rama de la IA que se dedica al estudio de aquellos agentes o programas de software que aprenden o evolucionan basados en su experiencia, para realizar una tarea determinada cada vez mejor. El objetivo principal es desarrollar técnicas que permitan a las computadoras aprender. Existen 4 formas distintas de aprender: mediante un aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi supervisado o aprendizaje por refuerzo [22].

En el caso particular de los 3 primeros tipos de algoritmos se diferencian en el conocimiento a priori que

se tiene en cada uno. Los dos extremos son el supervisado, donde se tiene conocimiento a priori de los datos, y el no supervisado, caracterizado por la ausencia de conocimiento a priori.

- Aprendizaje supervisado: Son los algoritmos más sencillos de realizar y comprender. En ellos se parte de un conocimiento a priori. El objetivo es, mediante unos datos de entrenamiento, deducir una función que haga lo mejor posible el mapeo entre unas entradas y una salida. Los datos de entrenamiento constan de tuplas (X, Y), siendo X las variables que predicen una determinada salida Y. La variable a predecir Y puede ser una variable cuantitativa para los casos de problemas de regresión o de tipo cualitativa para los casos de problemas de clasificación de un modelo de aprendizaje supervisado [23].

  Dentro de los algoritmos supervisados se pueden encontrar dos corrientes, los algoritmos de clasificación y los de regresión. Los algoritmos de clasificación se usan cuando el resultado deseado es una etiqueta discreta. Aquí tenemos clasificación binaria, solo se elige entre dos etiquetas y clasificación para múltiples etiquetas.

  La otra corriente son los algoritmos de regresión, los cuales son útiles para predecir valores que son continuos. Eso significa que la respuesta a su pregunta se representa mediante una cantidad que puede determinarse de manera flexible en función de las entradas del modelo en lugar de limitarse a un conjunto de posibles etiquetas. Consiste en un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente, las variables y un término independientes B. Este modelo puede ser expresado en forma de recta.
- Aprendizaje no supervisado: Se organizan los datos de alguna manera que se pueda describir su estructura. Esto puede significar agruparlos en clústeres o buscar diferentes maneras de examinar datos complejos para que parezcan más simples u organizados. Al contrario que en el aprendizaje supervisado, en este caso no existe conocimiento a priori. No se tienen tuplas (X,Y), simplemente se tiene la variable X. El objetivo del aprendizaje no supervisado es lograr crear un modelo de la estructura o distribución de los datos para aprender más sobre ellos. Sirve tanto para entender como para resumir un conjunto de datos. En términos generales, pueden ser agrupados en algoritmos de clustering y algoritmos de asociación.
- Aprendizaje semisupervisado: El aprendizaje semisupervisado se encuentra a medio camino entre el aprendizaje supervisado y el no supervisado. Ahora lo que tenemos son tanto datos etiquetados como datos no etiquetados, es decir, además de tener tuplas (X,Y), tenemos datos sólo de X de los que no sabemos su respuesta Y.
- Aprendizaje por refuerzo: En estos casos lo que tenemos son problemas no supervisados que sólo reciben realimentaciones o refuerzos (por ejemplo, gana o pierde). Se sustituye la información supervisada (Y) por información del tipo acción/reacción. El objetivo en el aprendizaje por refuerzo es aprender a mapear situaciones de acciones para maximizar una cierta función de recompensa. En estos problemas un agente aprende por prueba y error en un ambiente dinámico e incierto. En cada interacción el agente recibe como entrada un indicador de estado actual y selecciona una determinada acción que maximice una función de refuerzo o recompensa a largo plazo.

Elegir un algoritmo es un paso crítico en el proceso de aprendizaje automático, por lo que es importante que realmente se adapte al caso de uso del problema en cuestión [17]. Específicamente para la tarea clasificación de opiniones tenemos los algoritmos: máquina de soporte vectorial (SVM), las redes bayesianas (Naive Bayes), los árboles de decisión (Random Forest) y K-vecinos más cercanos (Knn). Para el desarrollo de nuestra investigación hemos utilizado los algoritmos de aprendizaje supervisado [24].

2.3 Trabajos relacionados.

En la actualidad algunos sistemas de AS se basan en técnicas de aprendizaje automático para el proceso de clasificación de sentimientos. Por ejemplo uno de los primeros trabajos de sistemas de AS que utiliza el enfoque de aprendizaje automático es el de [25], en estos trabajos se utilizan varios algoritmos de clasificación supervisados tales como Naive Bayes y máquina de soporte de vectores (SVM).

Propuestas, como la de Cruz [26] , han sido realizadas de forma flexible en cuanto a la aplicación de técnicas de clasificación de sentimientos basadas en aprendizaje automático. En este tipo de trabajos se puede aplicar cualquier método de clasificación sentimental.

En los trabajos de Min y Park [27] se enfocan en la combinación de dos algoritmos de aprendizaje automático. En sus investigaciones combinan SVM y reglas de árboles de decisión. En un trabajo de Chen [28] se ha utilizado otra técnica de aprendizaje automático para el AS llamada Condicional Random Fields (CRF).

Enfoques tales como [29], en donde se utilizan algoritmos basados en diccionario para llevar a cabo la

clasificación de sentimientos. Por último, y no menos importante el sistema presentado por [30] en el cual utilizan un servicio de terceros de AS llamado OpenDover. La utilización de este servicio no es gratuita y no está disponible su código fuente, por lo que se dificulta la obtención de dichas técnicas de AS empleadas ya que no están publicadas.

3. Materiales y métodos

Realizar AS en inglés es una tarea relativamente sencilla ya que existen paquetes que vienen con modelos preparados para calcular el sentimiento o polaridad de un nuevo texto. Sin embargo en español aún no se cuenta con muchas herramientas [31].

3.1 Propuesta

La metodología propuesta que se presenta en este artículo ha sido implementada como un caso de estudio en el siguiente orden: se investigó a cerca del estado del arte en el contexto de aprendizaje automático y AS, revisión de trabajos relacionados, obtención de un corpus etiquetado, regularización del corpus, implementación de técnicas para el pre proceso del texto hasta obtener el más óptimo, aplicación de dos algoritmos de aprendizaje automático supervisado para clasificación y análisis de los resultados [32].

3.2 Conjunto de datos

Se han utilizado varios conjuntos de datos (data set) en español para llevar a cabo la investigación. Este conjunto de datos fue recuperado del corpus TASS que es el Taller de Análisis de Sentimientos en español organizado cada año por la Sociedad Española del Procesado del Lenguaje Natural (SEPLN). Para conseguir el corpus nos pusimos en contacto con los organizadores vía mail y nos dieron acceso al uso de su repositorio de data set [33].

Dos conjuntos de datos principales para fines de aprendizaje se usaron: en general este corpus contiene 7.219 que son mensajes de Twitter escritos en español sobre personalidades conocidas en política, economía, comunicación o cultura. Algunos de estos archivos están enfocados en un tópico, por ejemplo, política o TV.

Los archivos del corpus están en formato XML y contienen miles de tweets en español etiquetados con su sentimiento o polaridad. Es decir, el texto estos categorizados si expresa algo en positivo, negativo o sentimiento neutral. En nuestro caso particular filtramos y solo utilizamos las polaridades positivas y negativas. Los temas son variados algunos mensajes son de política, fútbol, literatura o entretenimiento. En la figura 2 se muestra un extracto de la estructura de los archivos:

```
<tweet>
    <tweetid>756700943043928064</tweetid>
    <user>105199059</user>
    <content>Gracias a Dios cerrando con broche de oro!
    <date>Sat Jul 23 04:02:18 +0000 2016</date>
    <lang>es</lang>
    <sentiment>
        <polarity><value>P</value></polarity>
    </sentiment>
</tweet>
```

Figura 2. Extracto del corpus evaluado.

3.3 Preparando los datos

Una vez obtuvimos el conjunto de datos en XML los transformamos a formato CSV para su posterior lectura en Python. El texto en bruto es bastante desordenado, si analizamos el conjunto de datos, nos damos cuenta de que hay dentro del texto innumerables caracteres que sobran ya que no tienen una representación ni valía para nuestro trabajo por lo que se realizó una limpieza, en este caso se eliminaron utilizando una función de expresión regulares, que es un requisito previo para realizar cualquier tarea de PLN [34]. Con los datos limpios se hace necesario que el texto tenga sentido para el algoritmo, es decir necesitamos convertir el texto en a una representación numérica, es decir en un formato vectorial.

3.4 Librerías de Python

Se ha decidido trabajar con la herramienta de programación Python, el lenguaje de programación Python ofrece muchos beneficios para los que desean integrarse en el contexto de PLN y aprendizaje automático, pues posee una enorme cantidad de librerías que facilitan las tareas, entre ellas tenemos:

- **Sklearn:** Librería de aprendizaje automático que incluye muchos algoritmos de machine learning, aquí, estamos usando algunos de sus módulos como train_test_split, RandomForestClassifier y precision_score.
- **NumPy:** Librería con una variedad de módulos numéricos de Python que proporciona funciones matemáticas rápidas para los cálculos. Se utiliza para leer datos y llevarlos a arreglos y para poder manipularlos.
- **Pandas:** Es utilizada esta librería para leer y escribir en archivos. La manipulación de datos se puede hacer fácilmente con esta librería.

3.5 Preprocesar los datos

El algoritmo necesita algún tipo de vector de características para realizar la tarea de clasificación. La forma más sencilla de convertir un corpus a un formato vectorial, donde cada única palabra en un texto será representada por un número. Para ello, usamos una técnica

llamada vectorización o bolsa de palabras a través de la clase de Python llamada CountVectorizer [19]. En este paso se convirtió el texto en una matriz en la que cada palabra es una columna cuyo valor es el número de veces que dicha palabra aparece en el texto.

Antes de entrenar el modelo, tenemos que dividir el conjunto de datos en conjunto de datos utilizado en entrenamiento y prueba, esta tarea la realizamos utilizando el módulo de sklearn train_test_split.

3.6 Entrenamiento del algoritmo

Aquí es donde comenzamos a utilizar las técnicas de aprendizaje automático. En este paso se entrenaron los algoritmos de clasificación a través de los vectores de características generados en el preproceso.

Se requirió dos conjuntos de datos uno de entrenamiento y otro para pruebas. El primer conjunto es utilizado para que el algoritmo "aprenda" de las diversas características de los documentos y el segundo conjunto sirve para evaluar el rendimiento del modelo obtenido. Dicho modelo, permite clasificar nuevos documentos, en este caso como positivos o negativos. La idea es que los algoritmos puedan extraer información útil de los datos que le pasamos para luego poder hacer predicciones.

Aplicamos los algoritmos Naives Bayes, Máquina de soporte de vectores (SVM) y también Arboles aleatorios (Random Forest) [35].

El clasificador Naive Bayes usa el Teorema de Bayes, con un supuesto de independencia entre los predictores. En términos simples, un clasificador Bayesiano asume que la presencia de una característica particular en una clase no está relacionada con la presencia de cualquier otra característica.

SVM es un algoritmo de aprendizaje automático supervisado que se puede utilizar tanto para tareas de clasificación como de regresión. SVM realiza la clasificación al encontrar el hiper plano que diferencia las clases que trazamos en el espacio n-dimensional.

Random Forest [36] es un algoritmo que mezcla una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

3.7 Validación del modelo

Se han utilizado métricas que son de las más utilizadas hoy día para la evaluación de desempeño de modelos de clasificación, para nuestro caso usaremos las métricas en el modelo de clasificación que acabamos de entrenar, estas métricas son: matriz de confusión, medida de F1 y la precisión.

En la librería sklearn de Python podemos encontrar estos métodos para realizar las métricas mencionadas así que para el estudio hemos usado las herramientas classification_report, confusion_matrix y accuracy_score [37].

El puntaje de precisión (accuracy score) lo hemos utilizado para calcular la precisión del clasificador entrenado. Con la matriz de confusión pudimos tener una mejor idea de cómo está clasificando nuestro modelo, ya que nos proporcionó un conteo de los aciertos y errores de cada una de las clases por las que clasificamos. Pudimos comprobar si nuestro modelo está confundiéndose entre clases, y en qué medida.

En cada columna de la matriz de confusión tenemos el número de predicciones para cada clase realizadas por el modelo, y cada fila representa los valores reales por cada clase. Con lo cual los conteos quedan divididos en 4 clases, VP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos) y VN (ciertos negativos). En la figura 3 mostramos los elementos de la matriz de confusión [38].

		Predicción	
		Positivos	Negativos
Observación	Positivos	VP	FN
	Negativos	FP	VN

Figura 3. Elementos de la matriz de confusion.

- VP: Representa las predicciones correctas para la clase.
- FN: La predicción es negativa cuando realmente el valor tendría que ser positivo.
- FP: Son el número de falsos positivos, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo.
- VN: Son el número de verdaderos negativos, es decir, de predicciones correctas para la clase.

A través del método classification_report podemos crear un informe de texto que muestre las principales métricas de clasificación. Nos muestra un resumen en base a las metricas de: precision, recall, f1-score y support [39].

Los promedios reportados incluyen micro estos promedios se han obtenido promediando el total de positivos verdaderos, falsos negativos y falsos positivos, promedio macro promediando la media no ponderada por etiqueta, promedio ponderado promediando la media ponderada por soporte por etiqueta.

4. Resultados

Pudimos medir en el experimento el classification_report que nos brindó información de las métricas comunes de precisión, recall y f1-score para cada etiqueta de POS o NEG de las clases evaluadas con los distintos algoritmos (Naives Bayes, Random Forest, SVM) por lo observado los algoritmos de Random Forest y SVM presentan mejores resultados de precisión. Esto en gran medida a que siempre se ambos algoritmos son considerado mejores para la tarea de clasificación.

En la siguiente gráfica de la figura 4, podemos observar los resultados del classification_report:

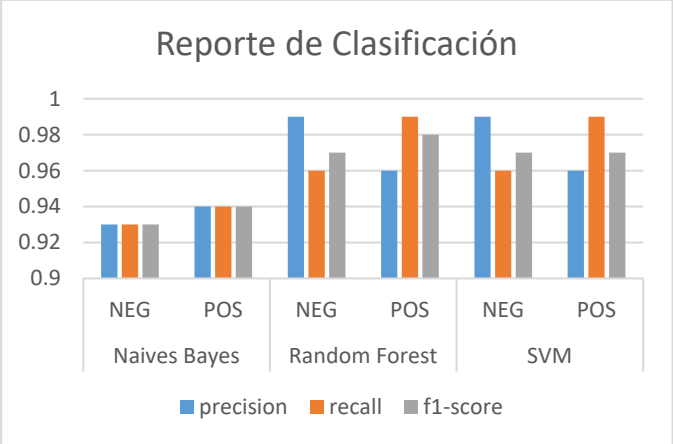

Figura 4. Reporte de clasificacion.

En cuanto a la precisión (accuracy) se refiere a que la hemos podido evaluar objetiva e individualmente en cada modelo. Dando como resultado que el algoritmo de Random Forest presenta un 0.9751 el mejor porcentaje obtenido de los tres. SVM segundo con 0.9735 y por último Naives Bayes con 0.9359, sin embargo, los tres están arriba del 93% lo que es bastante significativo y predice que los modelos están trabajando bien.

5. Discusión

Se han realizado diferentes experimentos. Se basan principalmente en el enfoque de aprendizaje automático, donde hemos combinado resultados de diferentes algoritmos de clasificación y también hemos podido utilizar métricas para ver su precisión.

Después de realizar los experimentos sobre el corpus de TASS en donde les aplicamos distintos algoritmos de aprendizaje automático para la tarea de clasificación de mensajes el rendimiento global obtenido es bastante positivo, ya que cada algoritmo mostro métricas superior al 93% en exactitud.

Aplicamos varios algoritmos de aprendizaje automático para clasificación, en conjunto los tres obtuvieron buenos resultados sin embargo para nuestro experimento nos resultó mejor Random Forest, lo que marca un precedente para otros trabajos similares.

La precisión es un tema importante en el AS. La precisión aún puede mejorarse, consideramos haciendo una selección más cuidadosa del corpus y el etiquetado de polaridad, puesto que en el corpus utilizado muy bien se ven textos valorados como positivos y que pudiesen ser lo contrario, mucho de esto por el cambio cultural o demás. Por lo tanto, este tipo de desafíos se pueden resolver utilizando enfoques innovadores.

El enfoque utilizado de bolsa de palabras produce una buena precisión en comparación con otros enfoques estudiados, pero requiere más esfuerzo computacional y es un método más lento.

6. Conclusiones y trabajos futuros

Se ha presentado una investigación que hace experimentos de AS sobre un corpus de datos que estaba etiquetado con polaridades positivas y negativas, aplicando diferentes algoritmos de aprendizaje automático supervisado y realizando tareas de clasificación. Nuestra meta fue analizar y validar a través de métricas cual algoritmo de los algoritmos presentados de clasificación era más eficiente para nuestros propósitos.

Para ello creamos una metodología que involucró el levantamiento del estado del arte, luego la selección y limpieza de los datos, el preprocesamiento, generación del modelo y su validación.

Consideramos que esta investigación fue positiva y eficiente en cuanto a la metodología propuesta y por otro lado se pudo ver las ventajas que ofrecen el aprendizaje automático en cuanto a la clasificación y procesamiento de texto no estructurado.

En cuanto a trabajos futuros, seguir encaminados en esta área de la IA pues nos parece muy interesante y proponer otros estudios con otro set de datos para ver el comportamiento de los algoritmos de clasificación.

AGRADECIMIENTO

Agradecemos a la Secretaria Nacional de Ciencia, Tecnología e Innovación (SENACYT), por el apoyo dado en el desarrollo de esta investigación y al Centro Regional de Panamá Oeste de la Universidad Tecnológica de Panamá (CRPO-UTP).

REFERENCIAS

[1] J. G. Montalvo, “Big Data & Data Science : situación actual y aplicaciones financieras,” 2014.

[2] M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4241–4251, 2013.

[3] T. Wilson and P. Hoffmann, "OpinionFinder: A system for subjectivity analysis," *Proc. hlt/emnlp Interact. Demonstr.*, no. October, pp. 34–35, 2005.

[4] B. Pang and L. Lee, *LR..references..Opinion Mining and Sentiment Analysis*, vol. 2, no. 1–2. 2008.

[5] N. Barrett, J. H. Weber-Jahnke, and V. Thai, "Engineering natural language processing solutions for structured information from clinical text: Extracting sentinel events from palliative care consult letters," *Stud. Health Technol. Inform.*, vol. 192, pp. 594–598, 2013.

[6] I. Peñalver-Martinez *et al.*, "Feature-based opinion mining through ontologies," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5995–6008, 2014.

[7] D. C. Moreno and M. V. Lombardo, "Ontología y Procesamiento de Lenguaje Natural," *KnE Eng.*, vol. 3, no. 1, p. 492, 2018.

[8] O. Araque, I. Corcuera, C. Román, C. A. Iglesisas, and J. F. Sánchez-Rada, "Aspect based Sentiment Analysis of Spanish Tweets," *Tass 2015*, pp. 29–34, 2015.

[9] G. a. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[10] J. Allan, "NLP for IR Natural Language Processing for Information Retrieval Notice to the reader," pp. 1–50, 2000.

[11] V. C. Pande and A. S. Khandelwal, "A Survey Of Different Text Mining Techniques," *IBMRD's J. Manag. Res.*, vol. 3, no. 1, pp. 125–133, 2014.

[12] C. Friedman, T. C. Rindflesch, and M. Corn, "Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine," *J. Biomed. Inform.*, vol. 46, no. 5, pp. 765–773, 2013.

[13] D. Cedeno-Moreno and M. Vargas-Lombardo, "An ontology-based knowledge methodology in the medical domain in the Latin america: The study case of republic of Panama," *Acta Inform. Medica*, vol. 26, no. 2, pp. 98–101, 2018.

[14] J. Villena-Román, S. Collada-Pérez, S. Lana-Serrano, and J. C. González-Cristóbal, "Método híbrido para categorización de texto basado en aprendizaje y reglas," *Proces. Leng. Nat.*, vol. 46, pp. 35–42, 2011.

[15] S. Ordoñez-Salinas and A. Gelbukh, "Information retrieval with a simplified conceptual graph-like representation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6437 LNAI, no. PART 1, pp. 92–104, 2010.

[16] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, 2014.

[17] J. Khairnar and M. Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification," *Int. J. Sci. Res. Publ.*, vol. 3, no. 6, pp. 1–6, 2013.

[18] C.-H. Lee and S.-H. Wang, "An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8954–8967, 2012.

[19] V. B. and B. M., "Analysis of Various Sentiment Classification Techniques," *Int. J. Comput. Appl.*, vol. 140, no. 3, pp. 22–27, 2016.

[20] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014.

[21] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[22] J. Brank and D. Mladenic, "Gold Standard Based Ontology Evaluation," *Work. Eval. Ontol. Web, EON*, 2006.

[23] A. Akusok, K.-M. Bjork, Y. Miche, and A. Lendasse, "High-Performance Extreme Learning Machines: A Complete Toolbox for Big Data Applications," *IEEE Access*, vol. 3, pp. 1011–1025, 2015.

[24] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[25] X. Zhu, S. Kiritchenko, and S. Mohammad, "NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets," no. SemEval, pp. 443–447, 2015.

[26] C. G. Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J. & Vallejo, "Long autonomy or long delay? The importance of domain in opinion mining.," *Expert Syst. Appl.*, vol. 40, pp. 3174–3184, 2013.

[27] J. C. Min, H.J., Park, "Identifying helpful reviews based on customer's mentions about experiences.," *Expert Syst. Appl.*, vol. 39, pp. 11830– 11838., 2012.

[28] F. Chen, L., Qi, L., Wang, "Comparison of feature-level learning methods for mining online consumer reviews.," *Expert Syst. Appl.*, vol. 39, pp. 9588–9601, 2012.

[29] J. Eirinaki, M., Pisal, S., & Singh, "Feature-based opinion mining and ranking.," *J. Comput. Syst. Sci.*, vol. 78(4), pp. 1175–1184, 2012.

[30] N. Kontopoulos, E., Berberidis, C., Dergiades, T. & Bassiliades, "Ontology-based sentiment analysis of twitter posts.," *Expert Syst. Appl.*, vol. 40, pp. 4065– 4074., 2013.

[31] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decis. Support Syst.*, vol. 48, no. September, pp. 354–368, 2010.

[32] T. Baldwin, P. Cook, B. Han, A. Harwood, S. Karunasekera, and M. Moshtaghi, *A Support Platform for Event Detection using Social Intelligence*. 2012.

[33] D. Vilares, Y. Doval, M. A. Alonso, and C. Gómez-Rodríguez, "LyS at TASS 2014: A Prototype for Extracting and Analysing Aspects from Spanish tweets," *Tass 2014*, 2014.

[34] a Holzinger, R. Geierhofer, F. Mödritscher, and R. Tatzl, "Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses," *J. Univers. Comput. Sci.*, vol. 14, no. 22, pp. 3781–3795, 2008.

[35] M. W. Berry, "Survey of Text Mining : Clustering, Classification, and Retrieval," *New York*, p. 262, 2004.

[36] G. Williams, "Hands-On Data Science with R Text Mining," no. January, 2016.

[37] I. Spasić, J. Livsey, J. A. Keane, and G. Nenadić, "Text mining of cancer-related information: Review of current status and future directions," *Int. J. Med. Inform.*, vol. 83, no. 9, pp. 605–623, 2014.

[38] N. Ur-Rahman and J. a. Harding, "Textual data mining for industrial knowledge management and text classification: A business oriented approach," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 4729–4739, 2012.

[39] M. Othman, H. Hassan, R. Moawad, and A. M. Idrees, "Using NLP Approach for Opinion Types Classifier," *J. Comput.*, vol. 11, no. 5, pp. 400–410, 2018.