ANÁLISIS AUTOMÁTICO DE GRANDES VOLÚMENES DE DATOS EN REDES SOCIALES MEDIANTE MINERÍA DE TEXTOS COMBINADO CON ALGORITMOS INTELIGENTES

Juan Pablo Tessore [1 2], Leonardo Esnaola [2], Claudia Russo [2 3], Hugo Ramón [2 3], Sabrina Pompei [2]

Instituto de Investigación y Transferencia en Tecnología (ITT)[4]
Comisión de Investigaciones Científicas (CIC)
Escuela de Tecnología (ET)
Universidad Nacional del Noroeste de la Provincia de Buenos Aires
(UNNOBA)

Sarmiento y Newbery, 236-4636945/44

{juanpablo.tessore, leonardo.esnaola, claudia.russo, hugo.ramon, sabrina.pompei}@itt.unnoba.edu.ar

RESUMEN

El presente trabajo propone construir un clasificador automático de opiniones, que permitirá realizar análisis automáticos a bajo costo del juicio de los consumidores acerca de productos o servicios. Dicho clasificador será entrenado a partir de los comentarios en lenguaje informal presente en redes sociales.

Para alcanzar el objetivo descripto, en primer lugar, se prevé construir una base de datos que reúna diversos fragmentos de texto en idioma español, incorporando los modismos propios de nuestra región.

En segundo lugar, a través de un proceso incremental de limpieza y normalización de cada fragmento de texto, que incluye actividades como la eliminación de hashtags, enlaces, emoticones, etc.; corrección ortográfica; etiquetado sintáctico (también conocido como "Part Of Speech Tagging", o simplemente "POS tagging"); desambiguación, entre otras.

Una vez realizada la recopilación y normalizado el contenido, se definirá un criterio de clasificación de dichos fragmentos, de manera de establecer clases que permitan agrupar los mismos según su afinidad, es decir a partir de características comunes.

Finalmente, a partir del diseño, desarrollo e implementación de un algoritmo inteligente se buscará determinar el grado de pertenencia a cada uno de los grupos definidos de cualquier texto arbitrario.

Palabras clave: Text mining, Big Data, Inteligencia artificial, Redes sociales.

---

1 Becario de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC)
2 Docente Investigador en el Instituto de Investigación y Transferencia en Tecnología (ITT) / Escuela de Tecnología / UNNOBA
3 Investigador Asociado Adjunto sin director a la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC)
4 Centro Asociado a la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC)

CONTEXTO

Esta línea de investigación forma parte del proyecto "Tecnología y Aplicaciones de Sistemas de Software: Calidad e Innovación en procesos, productos y servicios" aprobado por la Secretaría de Investigación, Desarrollo y Transferencia de la UNNOBA en el marco de la convocatoria a Subsidios de Investigación Bianuales (SIB2017). A su vez se enmarca en el contexto de un plan de trabajo aprobado por la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y por la Secretaría de Investigación de la UNNOBA en el marco de la convocatoria "Becas de Estudio Cofinanciadas 2015 CIC Universidades del interior bonaerense".

El proyecto se desarrolla en el Instituto de Investigación en Tecnologías y Transferencia (ITT) dependiente de la mencionada Secretaría, y se trabaja en conjunto con la Escuela de Tecnología de la UNNOBA.

El equipo está constituido por docentes e investigadores pertenecientes al ITT y a otros Institutos de Investigación, así como también, estudiantes de las carreras de Informática de la Escuela de Tecnología de la UNNOBA.

1. INTRODUCCIÓN

Desde que las primeras computadoras programables fueron concebidas, las personas se preguntaron si tendrían la capacidad de pensar, de aprender y de convertirse en "máquinas inteligentes".

El campo de la ciencia que se encarga de resolver este interrogante se denomina inteligencia artificial. Se trata de un área multidisciplinaria, que a través de ciencias como las ciencias de la computación, la matemática, la lógica y la filosofía, estudia la creación y diseño de sistemas capaces de resolver problemas cotidianos por sí mismos, utilizando como paradigma la inteligencia humana [1]. Para que una máquina pueda comportarse de manera inteligente debería ser capaz de resolver problemas de la manera en que lo hacen los humanos, es decir, en base a la experiencia y el conocimiento [2]. Esto implica que debería ser capaz de modificar su comportamiento en base a cuán precisos son los resultados obtenidos comparados con los esperados.

En este sentido podemos encontrar tres grandes grupos de algoritmos de *Machine Learning* [3]:

- Algoritmos supervisados: estos algoritmos utilizan un conjunto de datos de entrenamiento etiquetados (preclasificados), los cuales procesan para realizar predicciones sobre los mismos, corrigiendolas cuando son incorrectas. El proceso de entrenamiento continúa hasta que el modelo alcanza un nivel deseado de precisión.
- Algoritmos semi-supervisados: combinan tanto datos etiquetados como no etiquetados para generar una función deseada o clasificador. Este tipo de modelos deben aprender las estructuras para organizar los datos así como también realizar predicciones.
- Algoritmos no supervisados: El conjunto de datos no se encuentra etiquetado y no se tiene un resultado conocido. Por ello deben deducir las estructuras presentes en los datos de entrada, lo puede conseguir a través de un proceso matemático para reducir la redundancia sistemáticamente u organizando los datos por similitud.

Dentro de esta clasificación podemos además encontrar un gran número de algoritmos específicos con diferentes características para el tratamiento de los datos. Entre los más relevantes encontramos:

- *Deep Learning* (DL): consiste en la utilización de algoritmos para hacer representaciones abstractas de la información y facilitar el aprendizaje automático [4].
- *Active Learning* (AL): es un caso especial de aprendizaje semi-supervisado donde el algoritmo de aprendizaje puede interactuar con un usuario u otra fuente de información para obtener los resultados deseados [5].
- *Support Vector Machines* (SVM): busca la maximización de la distancia entre la recta o el plano y las muestras que se encuentran a un lado u otro. En el caso que las muestras no sean linealmente separables se utiliza una transformación llamada *kernel* [6] [7].

Una de las principales áreas de la inteligencia artificial es el procesamiento del lenguaje natural (PLN) o minería de textos. Esta área se encarga de desarrollar algoritmos que permitan extraer información relevante a partir de diversos contenidos en forma de texto. Con el auge de los contenidos sociales, la generación de este tipo de contenidos ha crecido en forma exponencial, esto último crea la oportunidad de aplicar algoritmos de minería de textos para extraer patrones significativos [8].

Según [9], de entre las técnicas de *Machine Learning* mencionadas, SVM es la más comúnmente utilizada para el análisis automático de textos, esto es así principalmente porque esta técnica es más apropiada para resolver problemas con una gran cantidad de dimensiones. Dentro de las tareas de minería de textos que pueden realizarse con esta técnica, podemos encontrar: clasificación de subjetividad; determinación de polaridad; resolución de ambigüedades; extracción de palabras de opinión y/o aspectos; etc.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La presente investigación se encuadra dentro de los ejes “Gestión de la innovación” y “Cómputo Ubicuo”. En ese sentido, los algoritmos basados en *Machine Learning* posibilitan, a través de su capacidad de aprendizaje y de su comportamiento inteligente de manera automática, el desarrollo de sistemas de cómputo ubicuo los cuales permiten, en última instancia, el monitoreo (y cambio) de la conducta humana y del ambiente donde esta se desarrolla.

Para la efectiva implementación de algoritmos de *Machine Learning*, se deberán abarcar las siguientes cuestiones:

- Obtención de un conjunto de datos, en formato de texto, suficientemente representativo para la problemática que se desea abordar.
- Pre procesamiento de las señales para lograr su normalización y adecuación.
- Determinación de categorías que permitan clasificar los textos recopilados según características comunes.
- Análisis y selección de los distintos algoritmos de minería de textos y *Machine Learnin*g que permitan clasificar los textos según las categorías definidas en el inciso anterior.

- Evaluación de fiabilidad y desempeño de las diferentes técnicas aplicadas.

3. RESULTADOS OBTENIDOS / ESPERADOS

Se espera que la presente línea de I/D permite adquirir conocimientos específicos sobre las diferentes técnicas de minería de textos y *Machine Learning*, con el propósito de desarrollar modelos capaces de predecir y clasificar los contenidos involucrados en la problemática que se intenta resolver, obteniendo un comportamiento inteligente de manera automática.

También se prevé la aplicación de *Machine Learning* y minería de textos en el análisis del texto redes sociales, con la finalidad de encontrar patrones dentro de esos datos que permitan predecir comportamientos futuros en ámbitos específicos.

Se espera como resultado final de esta línea de investigación, la creación de un producto transferible que permita llevar a cabo estudios sobre bienes y servicios, ofrecidos por personas y organizaciones pertenecientes a los sectores público y privado, detectando las opiniones manifestadas indirectamente por los comentarios de sus consumidores, sin la necesidad de destinar cuantiosos recursos a un análisis pormenorizado de los mismos.

Así mismo, se busca generar informes técnicos en base al trabajo realizado, en donde se registren los avances, el grado de implementación y los resultados obtenidos. Como así también difundir y transferir los resultados y logros alcanzados mediante la presentación y participación en diferentes congresos, jornadas y workshops de carácter nacional e internacional vinculados a la temática de estudio.

4. FORMACIÓN DE RECURSOS HUMANOS

En esta línea de I/D se han obtenido y se encuentran desarrollando actualmente una Beca de Estudio Cofinanciada otorgada por la Comisión de Investigaciones Científicas (CIC) y la UNNOBA. Asimismo se espera desarrollar una tesis doctoral y dos tesinas de grado, dirigidas por miembros de este proyecto.

5. BIBLIOGRAFÍA

[1] Assessment of the Commercial Applicability of Artificial Intelligence in Electronic Businesses. Thomas Kramer. Diplom.de. 2002.

[2] Data Classification Algorithms and Applications, Charu C. Aggarwal, CRC Press, 2015.

[3] Machine Learning An Algorithmic Perspective Second Edition, Stephen Marsland, CRC Press, 2015.

[4] A Deep Learning. Book in preparation for MIT Press. Bengio, Y., Goodfellow, I. and Courville, USA, 2015.

[5] Active Learning Literature Survey, Settles Burr, Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2014.

[6] A Tutorial on Support Vector Machines for Pattern Recognition, Christopher J.C. Burges, Kluwer Academic Publishers, 1998.

[7] Top 10 algorithms in data mining,

Xindong Wu et al. Knowledge and Information Systems 2008.

[8] Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python, Manohar Swamynathan, APRESS, 2017.

[9] A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Ravi Kumar et. al. IEEE Knowledge-Based Systems, 2015.