Propuesta de análisis de sentimientos de sitios web comunicacionales de Educación Superior. Un estudio del contexto regional.

Sonia I. Mariño, Silvina Podesta, Jaquelina E. Escalante

Departamento de Informática, Facultad de Ciencias Exactas y Naturales y Agrimensura,

Universidad Nacional del Nordeste, 9 de Julio 1449, Corrientes, Argentina

simarinio@yahoo.com, sipodesta@yahoo.com.ar, jaquelin_escalante@hotmail.com

Resumen

Uno de los retos de la Inteligencia Artificial del siglo XXI es reflexionar en torno a vínculos entre los sujetos –productores y consumidores de tecnologías, y éstas últimas. En un intento de explicar y comprender a los seres humanos ha surgido como tecnología de la Inteligencia Artificial el Análisis de Sentimientos, integrando teórica-metodológicamente conceptos y herramientas del Procesamiento de Lenguaje Natural y Aprendizaje de Maquinas. En el artículo se expone un caso de estudio preliminar del análisis de sentimiento aplicado a sitios web de comunicación institucional en espacios de educativos de la región.

Palabras Clave- Inteligencia Artificial, Análisis de Sentimientos, Educación Superior.

1. INTRODUCCIÓN

1.1 TIC en Educación Superior

La tecnología, ha sido tematizada como problema social en las últimas décadas, pasando a ocupar un lugar destacado en los medios de comunicación, los foros públicos y las agendas políticas. Con el intenso desarrollo tecnológico actual, se ha hecho especialmente evidente la estrecha dependencia de la economía, las instituciones y las formas de vida respecto de artefactos y procesos tecnológicos [1].

Con el mencionado cambio en las comprensiones públicas y académicas, entre finales de los años 60 y principios de los 70, se configuraron nuevos paradigmas en educación, como es el uso de las TIC; brindando herramientas para la adquisición del conocimiento mediado tecnológicamente [2, 6].

Para adaptarse a las necesidades de la sociedad actual, las instituciones de educación superior deben flexibilizarse y desarrollar vías de integración de las tecnologías de la información y la comunicación en los procesos de comunicación y formación. Paralelamente es necesario aplicar una nueva concepción de los alumnos-usuarios, así como cambios de rol en los profesores y en personal administrativo en relación con los sistemas de comunicación y con el diseño y la distribución de la información.

Lo expuesto implica modificaciones en las tendencias de comunicación hacia un modelo más flexible. Para entender estos procesos de cambio y sus efectos, así como las posibilidades reflejadas en los sistemas de Educación Superior que conllevan los cambios y avances tecnológicos, conviene situarse en los procesos de innovación

1.2 Análisis de Sentimientos

La Inteligencia Artificial, disciplina que formalmente nació en la década de 1950, ha tenido un vasto recorrido. A través de estos años han surgido y evolucionado las distintas tecnologías que la comprenden, dando lugar en numerosos casos a modelos híbridos que surgen de integrar algunas de sus áreas de conocimiento.

Actualmente dada la exposición de datos, información y conocimientos disponibles en la red de redes, uno de los retos de la IA en el siglo XXI gira en torno a capturar, analizar y reflexionar en torno al contenido semántico disponible. Es decir, el texto de los sitios web refleja cuestiones de género, edad, personalidades, sentimientos, entre otras variables de quienes los elaboran.

El Análisis de Sentimientos (AS), también se denomina como Extracción de Opiniones, Minería de Opiniones, Minería de Sentimientos o Análisis Subjetivo trata el estudio computacional de opiniones, sentimientos y emociones expresadas en textos [4, 5, 7, 9, 10, 11, 13].

El AS surge de integrar teórica-metodológicamente conceptos y herramientas del Procesamiento de Lenguaje Natural y Aprendizaje de Maquinas.

En la web se disponen de una diversidad de herramientas una de ellas es UClassify [8].

1.3 Enfoque del trabajo

A partir del marco teórico expuesto, el objetivo del presente trabajo es indagar en herramientas comprendidas en recuperación de información y análisis de sentimientos, como uno de los retos de la Inteligencia Artificial en el siglo XXI, y diseñar un estudio aplicado a un contexto de Educación Superior de la región NEA.

Se diferencia la recuperación de datos y la recuperación de información [3]. La primera intenta

recuperar todos los objetos que satisfacen claramente unas condiciones definidas expresadas mediante una expresión regula o una expresión del álgebra relacional [Baeza-Yates,99 citado en [3]).

La recuperación de información y el análisis de sentimientos han sido planteados como retos de la Inteligencia Artificial en el siglo XXI. Esto se debe a la explosión de datos y su disponibilidad en la web.

En [12] se analizan los sentimientos en un entorno de aprendizaje El trabajo involucra el tratamiento preliminar de análisis de sentimiento. Lo expuesto se justifica en la hipótesis que el género y el estado anímico -entre otras características de los sujetos redactores- implícitamente se refleja en el contenido desplegado en los sitios web institucionales.

Además, cuestiones de género se tratan en numerosos proyectos y estudios vigentes, ilustrando las relaciones de estos aspectos sociales y las TIC, enmarcados en la sociedad del conocimiento.

Por ello se aplicará una herramienta disponible en la web que permitiría realizar un análisis de género y estado emocional del texto disponible en los sitios web de comunicación institucional dependientes de una Universidad. Cabe aclarar que el texto desplegado en los sitios web es elaborado por humanos.

II. Metodología

En el trabajo se aplicó un método integrado por las siguientes fases:

- Selección de herramientas en línea orientadas al análisis de sentimientos.

En la web se localizan una diversidad de herramientas. En [7] se presenta un análisis de varias de ellas.

Particularmente, se optaron por distintos clasificadores provistos por la herramienta uClasify. Como inconveniente se explicita que funciona como una caja negra. Su parametrización implica la codificación, tema propuesto a futuro.

uClasify, brinda una diversidad de clasificadores que analizan en particular el texto seleccionado. Cabe aclarar que se eligieron algunos de los cuales se explicitan que NO trabajan con texto en inglés; dado que el contenido de las web seleccionadas se encuentra en idioma español.

Las herramientas de análisis de sentimientos, como otras comprendidas en inteligencia computacional, funcionan en dos modalidades entrenamiento y clasificación. Para la obtención de los resultados expuestos en el presente documento, se utilizaron en el segundo modo.

La herramienta implementa en su núcleo un clasificador Naïve Bayes multinomial, descripto en Mitchell (1997), fundamentado en el teorema de Bayes. Incluye un par de pasos que mejora aún más la clasificación (NB complementaria híbrida, la normalización de la clase). Como resultado de las clasificaciones, genera las probabilidades -en el rango [0-1]- de un documento perteneciente a cada clase. El resultado permite establecer un umbral para las clasificaciones. A fin de ejemplificar las clasificaciones de más del 90% se consideran spam. El uso de este modelo también permite que sea muy escalable en términos de tiempo de CPU para la clasificación.

- Aplicación de técnicas para el relevamiento de los datos

Los datos analizados por los clasificadores provinieron de los contenidos de sitios web institucionales de una Universidad Argentina pública. Se recopilaron las direcciones URL. Se incluyeron en las herramientas de clasificación y se obtuvieron resultados.

- Aplicación de técnicas para el análisis de los datos

Las herramientas de análisis de sentimientos seleccionadas, brindan datos porcentuales que pertenecen a distintas categorías. Se utilizan técnicas cualitativas a fin de explorar que representan estos valores situados en el contexto de aplicación. Cabe aclarar que el análisis de los resultados permitió elaborar inferencias preliminares, las que podrían corroborarse con datos reales en un futuro. Es decir, en el estudio se recurrió a técnicas interpretativas a fin de analizar y reconstruir la información producida por el clasificador.

Se obtuvo un documento con información de caracter:

- Interpretativa, que explicita una mirada retrospectiva –vinculada al contenido disponible en cada una de las webs institucionales- y
- Proyectiva, –planteada en una futura corroboración con datos reales con la finalidad de facilitar una lectura más sistemática del conocimiento plasmado como contenidos en los web de comunicación institucional.

Muestra

La técnica de muestreo aplicada se basó en el muestreo aleatorio simple e intencional o de conveniencia. Se elaboró un listado de las distintas dependencias de la Universidad. Finalmente. La muestra se conformó con el sitio web institucional de la Universidad y aquellos pertenecientes a las Facultades.

III. Resultados

Uno de los retos de la Inteligencia Artificial a través de diversos algoritmos computacionales es identificar, extraer y analizar la información disponible en diversos medios de comunicación del siglo XXI como son las redes sociales y los sitios web.

Los espacios de Educación Superior se constituyen en un riquísimo origen de datos para aplicar tecnologías de la IA, dado que permiten elaborar inferencias y a partir de allí mejorar el posicionamiento institucional, ya sea para captar nuevos sujetos o favorecer su permanencia en estos espacios de formación.

Por lo expuesto, a continuación se presentan distintas tablas que concentran los valores porcentuales derivados de aplicar diferentes clasificadores automáticos de texto, que utilizando la semántica implícita en los contenidos web proponen el análisis de sentimientos contemplando criterios como el género y los sentimientos de aquellos sujetos que elaboran noticias.

Las herramientas de procesamiento de textos seleccionados se encuentran comprendidas en la denominada clasificación binaria de la actitud de un texto. En el estudio, estas categorías pueden diferenciarse en femenino y masculino (Tabla 1) y en positiva o negativa (Tabla 2). Según el caso, también puede existir el neutro.

Lo expuesto precedentemente indica que el análisis de texto permitiría inferir si éste se redacta por una persona del género femenino, masculino o no es posible diferenciar, y si se encontraba ante una actitud positiva, negativa o neutra.

Cabe aclarar que estudios enfocados en cuestiones de edad u otras variables implicarían una tarea más compleja siendo viable la multiclasificación de un texto según el grado de polaridad de la actitud

En las Tablas 1 y 2 sintetizan la clasificación binaria de textos en idioma español utilizando como método de aprendizaje el algoritmo Naive Bayes.

En la Tabla 1 se expresa la actualización de noticias por sitio web según género. Se detecta que en el 23,08% de los casos analizados las noticias se actualizan de igual manera por varones o mujeres. Por otro lado se diferencia un alto porcentaje de elaboración de contenidos por mujeres que representa el 53,85%, siendo un 23,08% restante realizado por varones; como se observa en el Grafico 1.

Tabla 1.
Tipología por género en sitios web comunicacionales

Sitios pertenecientes a	***Mujeres***	***Varones***
Educa1	48	52
Educa2	9	91
Educa3	61	39
Educa4	42	58
Educa5	50	50
Educa6	56	44
Educa7	70	30
Educa8	94	6
Educa9	92	8
Educa10	67	33
Educa11	97	3
Educa12	50	50
Educa13	50	50

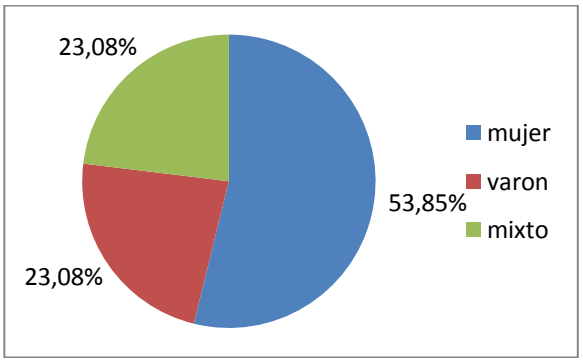

Gráfico 1: Tipología por genero de actualización de noticias

La Tabla 2 resume la *Tipología por sentimientos* estimada por el clasificador que determina si el contenido del sitio analizado es positivo o negativo. La herramienta uCclasify es adecuada para analizar textos cortos y largos (tweets, estados de Facebook, blogs, comentarios de productos, entre otros). La clasificación se basa en un modelo entrenado con 2,8 millones de documentos con datos provistos por Twitter, Amazon opiniones y críticas de películas.

Según se observa en la Tabla 2 y en el grafico 2, el nivel de contenido se expresa mayoritariamente en positivo con un valor del 84,62% frente al 7,69% que indica cierta tendencia a expresiones negativas. Un análisis en profundidad en la Tabla 2 permite distinguir que unos de los sitios (Educa2), afecta a las estimaciones de la herramienta.

Además, el clasificador determinó que el sitio denominado como Educa5 presenta en igual porcentaje

expresiones positivas y negativas representando el 7,69% restante de los resultados obtenidos (Grafico 2).

Tabla 2.
Tipología por sentimientos de los sitios web analizados

Sitios pertenecientes a	Positivo	Negativo
Educa1	100	0
Educa2	82	18
Educa3	100	0
Educa4	100	0
Educa5	50	50
Educa6	100	0
Educa7	100	0
Educa8	100	0
Educa9	100	0
Educa10	100	0
Educa11	100	0
Educa12	100	0
Educa13	100	0

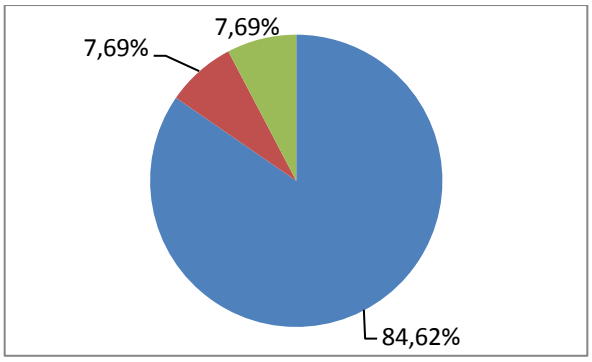

Gráfico 2: Aceptación por contenido presentado en los sitios web

IV Conclusiones

Se presentó un estudio interdisciplinar con intención de vincular distintos campos de conocimientos que involucran las tecnologías y las ciencias sociales. El primero representado por las Tecnologías de la Información que conforman una herramienta de la sociedad del conocimiento, y el segundo abordando estudios preliminares de género y emociones de quienes son responsables de transmitir información a través de sitios web institucionales..

La indagación se realizó utilizando métodos de aprendizaje automático (Machine Learning), en particular se optó por el algoritmo Naive Bayes, aplicado a la recuperación de textos disponibles en sitios web institucionales de una Universidad Argentina. El estudio se situó en un análisis de sentimientos preliminar, entendiéndose como uno de los retos de la Inteligencia Artificial en el siglo XXI.

Los resultados obtenidos por estos métodos automáticos se confrontarán con la realidad, de modo que una pesquisa determine la proximidad de los porcentajes estimados por la herramienta de clasificación con respecto a los valores reales medibles en los sujetos responsables de actualizar la información disponible en los mencionados sitios web institucionales.

También, se podría plantear determinar ciertos aspectos de un texto y sus sentimientos asociados atendiendo a otras variables.

A fin de profundizar en el análisis desde una mirada psicológica se podría trabajar interdisciplinariamente con un especialista en comportamientos humanos, quien aportaría en la definición de características de género y emoción que incidirían en la redacción de noticias o contenidos disponibles en la web de espacios de Educación Superior. Éstos estudios preliminares podrían corroborarse utilizando alguna herramienta de análisis de sentimientos.

Desde una mirada computacional, se propone:

- Analizar el comportamiento de distintos algoritmos o herramientas de clasificación utilizando el contenido tratado en este documento, así como incluir otros atributos atinentes a los contenidistas o redactores, a fin de proponer soluciones que podrían elevarse a los responsables institucionales con miras a la mejora de la imagen institucional.
- Aplicar técnicas de análisis de sentimientos previo al despliegue de cada sitio web en Internet y así evaluar la calidad de los datos utilizados para comunicar la información. Lo expuesto permitiría diseñar estrategias institucionales orientadas a mejorar la reputación organizacional y lograr una comunicación asertiva hacia el potencial auditorio.
- Avanzar en la definición de una interfaz de usuario para la generación de esta información de apoyo a la toma de decisiones destinada a los gestores universitarios a fin de asegurar experiencias de usuarios satisfactorias en los consumidores de información institucional.

Para finalizar se cita a Aristóteles quien sostuvo: “El problema de una emoción no es sentirla, sino saber cómo usarla”.

Referencias

[1] J. A. López Cerezo, J. L. Luján, Filosofía de la Tecnología *Revista Internacional de Filosofía*. *Tecnos* Vol. XVII/3 1998. Pág. 17

[2] B. Fainholc, *La interactividad en la educación a distancia*. 1999, 1ra Edición, Bs As. Paidós

[3] J. A. Olivas Varela, Las técnicas de Soft-Computing en la recuperación de información, 2011. Primer Seminario Doctoral en Aplicaciones y Transferencia de la Inteligencia Computacional (SEMATICA 2011) Disponible en http://eventos.citius.usc.es/sematica2011/PDFs/traspas_JAngelOlivas.pdf

[4] J. A. Olivas Varela, Algunos retos para la IA del Siglo XXI. Seminario de Posgrado, Misiones, 2016.

[5] B. Pang & L. Lee, Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr*., 2008, 2(1-2):1–135.

[6] R. Palomo López, J. Ruiz Palmero y J. Sánchez Rodríguez. Las TIC como agentes de innovación educativa. Junta de Andalucía. [En Líne] Disponible: http://www.juntadeandalucia.es/averroes/publicaciones/nntt/TIC_como_agentes_innovacion.pdf, 2006.

[7] J Serrano-Guerrero J. A. Olivas, F P. Romero, E. Herrera-Viedma Sentiment analysis: A review and comparative analysis of web services, *Information Sciences* 311 (2015) 18–38

[8] uClassify, https://www.uClassify.com/

[9] R. Valitutti, "WordNet-Affect: an Affective Extension of WordNet", en In Proceedings of the 4th International Conference on Language Resources and Evaluation, [En línea]. Disponible en: http://wndomains.fbk.eu/publications/lrec2004.pdf, 2004

[10] A. Esuli & F. Sebastiani "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining". [En línea]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.380.8135&rep=rep1&type=pdf [Accedido: 26-oct- 2016].

[11] M. M. Bradley & P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings", [En línea]. Disponible en: http://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf, Citeseer, 1999, [Accedido: 26-oct- 2016]

[12] L. Aballay, S. Aciar & E. Reategui "Propuesta de un Método para Detección de Emociones en E-Learning" *44 JAIIO - ASAI 2015,* [En línea]. Disponible en: http://44jaiio.sadio.org.ar/sites/default/files/asai121-128.pdf. 2015. [Accedido: 26-oct- 2016]

[13] R. Hernandez Petlachi & X. Li, "Análisis de sentimiento sobre textos en Español basado en aproximaciones semánticas con reglas lingüísticas". *Sociedad Española para el Procesamiento del Lenguaje Natural* (SEPLN). [En línea]. Disponible en: http://www.sepln.org/workshops/tass/2014/papers/4.CINVESTAV_IPN.pdf. 2014. [Accedido: 26-oct- 2016]