



NUEVAS REALIDADES PARA LA EDUCACIÓN EN INGENIERÍA:
CURRÍCULO, TECNOLOGÍA, MEDIO AMBIENTE Y DESARROLLO

13 - 16
DE SEPTIEMBRE

2022

CARTAGENA DE INDIAS,
COLOMBIA



Comparación de métodos de análisis de sentimientos en comunidades de habla hispana

Luis Gabriel Moreno Sandoval, Alexandra Pomares Quimbaya, Camilo Eduardo Cruz Gutiérrez, Johan Ferney García Pachón, David Felipe Vanegas Ramírez

Pontificia Universidad Javeriana
Bogotá, Colombia

Resumen

Uno de los fenómenos de estudio que ha motivado a sociólogos, antropólogos, psicólogos y, en general, a investigadores interesados en las interacciones humanas y los alcances de la comunicación lo constituye el Análisis de Subjetividad Colectiva. Las redes sociales son actualmente la principal plataforma para “escuchar” modos de pensar, actuar y sentir (PAS) entre usuarios en la Web con una infinidad de retos computacionales para determinar información útil. Por ello, analizar el sentimiento como un aspecto del lenguaje usado para expresar los estados mentales y emocionales o “estados privados”, que no pueden ser directamente observados, representa un aporte significativo para analizar la subjetividad colectiva. El presente artículo compara el desempeño de modelos de aprendizaje de máquina clásico, aprendizaje profundo y modelos basados en lexicones con el fin de establecer una metodología para analizar los PAS de las comunidades de habla hispana que puedan contribuir al análisis de subjetividad colectiva. Con este propósito, se utiliza un corpus de Twitter proveniente de la competencia TASS2019 para el idioma español. El enfoque de aprendizaje profundo obtuvo los mejores desempeños con los artículos de los principales periódicos en los 32 departamentos de Colombia.

Palabras clave: análisis de sentimientos; Twitter; habla hispana

Abstract

One of the study phenomena that has motivated sociologists, anthropologists, psychologists, and, in general, interested researchers in human interactions and the scope of communication is Collective Subjectivity Analysis. Social networks are currently the primary platform for "listening" to

thinking, acting, and feeling (TAF) modes between users on the web, with many computational challenges to determine useful information. Thus, analyzing sentiment as a language aspect used to express mental and emotional states or "private states," which cannot be directly observed, represents a significant contribution to analyzing collective subjectivity. This paper compares the performances of classical machine learning, Deep Learning, and lexicon-based models to establish a methodology for analyzing the TAF of Spanish-speaking communities that can contribute to the analysis of collective subjectivity. For this purpose, the paper used a Twitter corpus from the TASS2019 competition for the Spanish language. Deep Learning approach obtained the best performance with articles from the leading newspapers in the 32 Colombia states.

Keywords: *sentiment analysis; Twitter; spanish language*

1. Introducción

En los últimos años, las redes sociales han cambiado la relación entre las personas y el mundo que las rodea. Hoy en día se puede hablar, compartir, discutir, intercambiar información y tener interacciones similares a las del mundo real a través de foros web, blogs, páginas de fans, etc. (Roales, 2014). Como consecuencia, cada día se sube a Internet una enorme cantidad de datos estructurados y no estructurados. Esto ha provocado un gran interés en múltiples comunidades viéndolo como oportunidad para extraer y analizar las opiniones de los usuarios compartidos en las redes sociales y utilizarlo comercialmente. El Análisis de Sentimiento (AS) se centra en determinar la polaridad de un documento, frase o característica (Molina et al., 2013), de allí que se adapte directamente a la necesidad expuesta en el presente documento. Para abordar esta tarea, proponemos 3 métodos de clasificación: Aprendizaje de máquina clásico, Lexicones y Deep Learning.

Uno de los mayores retos de AS radica en la importancia de los dialectos, palabras coloquiales y el contexto de cada región (Utitiq et al., 2020). En consecuencia, destacamos el uso y entrenamiento de embebimiento de palabras específicas del español colombiano utilizando artículos de noticias de cada uno de los departamentos de Colombia. Además, se comparan con embebimientos genéricos pre-entrenados que fueron construidos con base en los datos de Wikipedia. En consecuencia, este documento se centra en la clasificación de sentimientos basada en un corpus de redes sociales en español. El objetivo del documento es comparar los tres métodos de clasificación propuestos y determinar cuál es el mejor y por qué.

2. Estado del arte

A través de la literatura consultada, el AS se ha abordado principalmente mediante dos enfoques: basado en lexicones y basado en el Machine Learning (ML). Recientemente se ha llevado a cabo un estudio utilizando el enfoque de lexicones (Moreno et al., 2017). En este estudio se proponen un nuevo conjunto de datos denominado CSL (Combined Spanish Lexicon), un enfoque que nace de una estrategia de bolsa de palabras ponderada y de la combinación de un conjunto de seis lexicones en español: iSOL, Elh Polar, SEL SLS, MI-SentiCon y MS, obteniendo un 62,38% de exactitud. En Molina et al. (2013), toman un conjunto de datos de lexicones en español formado por:



SWN, iSOL y BLEL para realizar varios experimentos utilizando diferentes enfoques no supervisados, el objetivo fue comparar diferentes metodologías sobre un corpus de críticas de películas y resolver el problema de la clasificación de la polaridad en español.

Por otro lado, el enfoque basado en ML sobre AS es el que ha tenido mejores resultados en comparación con la metodología basada en lexicones. Harjule et al. (2020) hacen uso de varios enfoques, entre los que destaca el Multinomial Naive Bayes. Este se basa en la suposición de que la aparición de diferentes palabras es independiente entre sí. Junto a este, se hace uso de la Support Vector Machines (SVM), un clasificador geométrico y no ignora las independientes. En sus experimentos, el SVM obtuvo una exactitud del 68,9%, mientras que el Multinomial Naive Bayes un 63,8%. También se afirma que cuanto mayor es el conjunto de datos (para ML), mejor es la exactitud, por lo tanto, los algoritmos de ML funcionan mejor que los algoritmos basados en lexicones, aunque requieren un conjunto de datos mayor para obtener mejores resultados.

Profundizando en el ML, los enfoques de Deep Learning (DL) son actualmente los más utilizados (Osorio et al., 2021) y están teniendo resultados significativos en comparación con los métodos de ML (Sun et al., 2017) (Nankani et al., 2020) (Osorio et al., 2021). En Kamyab et al. (2021) utilizan Embebimiento de palabras (EP) GloVe, Convolutional Neural Network (CNN) y Bilateral Long Short Term Memory (Bi-LSTM) en cuatro conjuntos de datos obteniendo un 87,12%, 94,01%, 91,94% y 94,53% respectivamente. Con base en ello, se evidencia que los resultados en AS difieren significativamente entre el inglés y el español, como afirman Osorio et al. (2021).

En particular para el español, en Chiruzzo et al. (2020) utilizaron EP y compararon enfoques ML (SVM) y DL (CNN y LSTM) para evaluar el sentimiento de las publicaciones en redes sociales en el conjunto de datos TASS2018 (15.893 tweets). Su mejor modelo fue un SVM con una exactitud del 56,2%, entre un rango del 49% y 60% para los restantes modelos. Finalmente, en Ochoa & Ari (2018) realizaron experimentos con tres tipos de EP (GloVe, Word2vec y Fasttext), Redes Neuronales Recurrentes (RNR) y CNN (también utilizándolas por separado) con el TASS. Obteniendo un 60% de exactitud como mejor resultado en el experimento RNN + CNN.

3. Metodología

Esta sección describe el preprocesamiento aplicado al corpus, las características de este y, por último, los tres métodos propuestos para la tarea de AS.

Preprocesamiento de datos

Los textos de las redes sociales no están naturalmente bien escritos, sobre todo en redes sociales como Twitter que tienen un número máximo de caracteres. Esto obliga a los usuarios a ajustar la estructura de sus textos; además que normalmente las personas no se detienen a corregir lo que escriben. Por ello, la preparación de los datos es un paso crucial para esta investigación.

Al principio, hacemos una preparación de datos llamada "*base*" que consiste en: convertir todas las letras en minúsculas; luego se remueven los nombres de usuario/menciones (@) y hashtags (\#);



paso seguido se eliminan todos los enlaces; después, se eliminan todos los caracteres que no sean letras; finalmente, eliminan las "Stop words" en español. En este orden de ideas, se realizan a continuación 4 preparaciones de datos diferentes para ver la incidencia de cada una de ellas en los modelos:

- **En bruto:** no hay más modificaciones aparte de la "base" descrita antes.
- **Lematizado:** Lematización de datos, sustituir cada palabra por su forma base (jugaro → jugar).
- **Normalizado:** Quitar las tildes (á → a).
- **Lematizado y normalizado:** Combinación de los dos últimos puntos. Dado que el acento puede cambiar el lema de la palabra, la lematización se hace primero.

Conjunto de datos

En esta sección se describen los conjuntos de datos utilizados para cada uno de los modelos.

Conjunto de datos por embebimiento de palabras

Este conjunto de datos está compuesto por artículos de noticias que fueron recopilados para los 32 departamentos de Colombia. En total son 740.065 artículos que se guardaron en la base de datos MongoDB. El proceso de recopilación de los periódicos se llevó a cabo durante 2 años. Los artículos se dividieron por categorías globalizadas, lo que significa que había varias categorías que eran iguales, pero tenían nombres diferentes. Para cada una de ellas, se estableció un único nombre y fue añadido como etiqueta en el conjunto de datos. Los resultados pueden verse en la tabla 1:

Tabla 1. Artículos de periódicos por categoría.

Categoría	Canti- dad	Catego- ría	Canti- dad	Categoría	Canti- dad
Deportes	122.752	Social	177.912	Entretenimiento	22.879
Economía	64.328	Regional	16.046	Medio ambiente	2.945
Política	61.897	Colombia	48.483	Opinión	21.132
Departmental	24.848	Mundo	24.345	Tecnología	10.579
Judicial	36.417	Salud	14.031	Ciencias	5.152
Cultural	84.836	Negocios	339	Mascotas	1.144

Modelando el conjunto de datos de entrenamiento

El Workshop on SA for the Spanish language (TASS) es un taller anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), proporcionando un conjunto de datos etiquetados. El conjunto de datos TASS 2019 fue utilizado (Villena et al, 2013), el cual contiene un tweet y sus correspondientes etiquetas de sentimientos: positivo (5.706 datos), negativo (5.265 datos) y neutro (2.013 datos) para 12.894 tweets en total.

El proceso general se resume en la figura 1. Primero, se recopilaron artículos de noticias para crear el conjunto de datos, después, se realizó el procesamiento y se guardaron los EP generadas. Una vez realizado el entrenamiento de las EP, se utilizaron los datos de las redes sociales para entrenar los tres métodos propuestos (Lexicons, ML y DL) y, por último, se almacenaron los resultados.



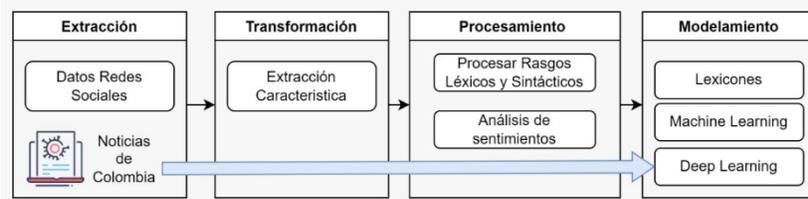


Figura 1. Proceso completo.

Lexicones

Para la evaluación del enfoque por lexicones, se buscaron conjuntos de lexicones en español, sin embargo, los recursos disponibles para este idioma son muy limitados. En consecuencia, sólo se dispone de un total de 4 conjuntos de datos de lexicones: ISOL (Secretaría de Estado de Investigación Desarrollo e Innovación, 2011)], ElhPolar (Moreno et al, 2017), Sentiment Spanish Lexicon (SSL) (Perez et al, n,d,) and ML-SentiCon (MS) (Data Science Lab, n,d,). Además, se propuso un quinto conjunto de datos denominado - Final Lexicon (FL) - basado en la combinación de los 4 anteriores. Los lexicones se distribuyeron de la siguiente manera:

Tabla 2. Estadísticas descriptivas de los lexicones.

	iSOL	ElhPo- lar	SSL	ML	FL
# Pos Words	2509	5302	1332	955	8182
#Neg Words	5626	1897	2511	1300	12199
#Neu_Words	0	0	0	9287	0
Polarity range	-1, 1	-1, 1	-1, 1	-1, 1	-1, 1

Una vez identificados los conjuntos de lexicones, se probó el rendimiento individual de cada uno. La prueba consistió en utilizar los tweets descritos en la tabla 2 para validar cuántos son calificados correcta e incorrectamente, donde la clasificación total de cada tweet se calculó como:

$$Sentimiento = \frac{\sum P(w)}{t}$$

Donde:

- w es una palabra dentro de una publicación en los medios sociales,
- t es el número total de palabras de la publicación en las redes sociales,
- P es la etiqueta de sentimiento de w en el conjunto de datos de los lexicones.

Cuando una palabra no pertenece al grupo de palabras del conjunto de datos de los lexicones, la palabra se ignora y se resta 1 del número total de palabras del tweet. Allí se calculan la puntuación $f1$, la precisión, la exhaustividad y la exactitud, que se utilizarán como métricas de rendimiento.

Aprendizaje de máquina clásico

Para los modelos de aprendizaje de máquina clásico, la biblioteca Scikit-Learn (Pedregosa et al, 2011) se utilizó para distribuir los textos con diferentes tipos de clasificadores: MultinomialNB,



SGDClassifier, BernoulliNB y KNeighborsClassifier. Utilizando las herramientas proporcionadas por estos cuatro, el objetivo es evaluar cuál es el mejor enfoque de ML teniendo en cuenta las métricas: f1-score, precisión y exhaustividad, Ahora, para el entrenamiento de cada uno de los modelos realizamos los siguientes pasos:

1. Cada uno de los textos se convierte en una serie de tokens numéricos mediante el tokenizador Tweet-Tokenizer incluido en la biblioteca NLTK (Bird et al, 2009).
2. Se utiliza CountVectorizer y TfidfTransformer con el objetivo de reducir el impacto de los tokens que ocurren con mucha frecuencia y que pueden ser menos informativos.
3. Se utilizan los 4 clasificadores. Este tipo de modelado de clasificación predictiva tiene como tarea la aproximación de una función de mapeo con variables de entrada a variables de salida discretas que en este caso serían: positivas, neutras y negativas.

Hay que tener en cuenta que previo a estos pasos, los textos deben ser procesados como se explica en la sección preprocesamiento de datos para evitar datos no deseados en el entrenamiento del modelo, Esto permite tener una mayor exactitud, precisión y puntuación F1.

Deep Learning con embebimiento de palabras

Para los modelos de DL, se utiliza una Red Neuronal recurrente LSTM (Long Short-Term Memory), Según Graves & Schmidhuber (2005), este tipo de redes tienen mejores capacidades para aprender las dependencias a largo plazo en los datos temporales utilizando células y puertas de memoria (Chandra & Krishna, 2021). Además, se combinan con EP previamente entrenadas (modelos Gensim, Bert y FastText) para obtener mejores resultados. El enfoque de este estudio es el siguiente, como se ve en la **Figura 2**:

- Se crea un codificador de texto con una capa TextVectorization, que permite la vectorización de un corpus de texto, convirtiendo cada texto en una secuencia de enteros.
- Se utiliza pad_sequences para asegurar todos los tokens tengan la misma longitud para poder ser entregadas a la red LSTM.
- Los datos se procesan en una capa de EP y se asigna un vector a cada uno de los codificadores que se generaron en el primer paso.
- Se procesan los datos en una red LSTM para almacenar la información sin que se produzca un desbordamiento en la red, lo que genera estabilidad.
- Finalmente, pasa por una capa lineal, que mapea la salida y se encarga de convertir la representación vectorial en un único *logit* como salida de la clasificación,



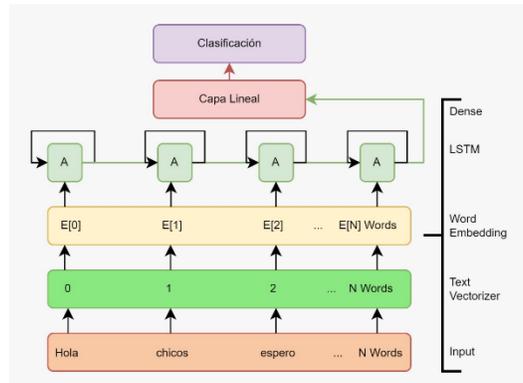


Figura 2. Arquitectura de Deep learning

Tras entrenar el modelo, se serializa y almacena para posteriormente evaluar el enfoque utilizando datos que el modelo no ha visto antes (10% del conjunto de datos). Ahora bien, se puede afirmar que a medida que aumentan las epochs, la exactitud también aumenta, lo que quizás implique un overfitting en el modelo. Teniendo en cuenta esto, se utilizaron dos técnicas para reducir el overfitting: la primera es una capa de abandono con un valor de 0,1 y la segunda es determinar la mejor cantidad de epochs con un punto de inflexión donde se obtiene una buena exactitud sin dejar de lado el incremento del overfitting,

Entrenamiento de embebimiento de palabras

Para la evaluación de la AS en ML, se propuso el uso de EP locales para capturar mejor el contexto de las palabras en español; para su entrenamiento se utilizó un corpus construido a partir de artículos de los periódicos mencionados (740,065 artículos). Con ellos, fueron entrenados varios modelos con cada uno de los siguientes elementos:

- FastText: Biblioteca para el aprendizaje eficiente de representaciones de palabras y clasificación de frases que construye vectores de palabras en función de un texto dado, con el fin de captar la información oculta de una lengua, como las analogías o la semántica (Facebook Inc, 2022). Se utilizó el algoritmo skipgram para el entrenamiento, dado que sus resultados son superiores a los del CBOW (van Paridon & Thompson, 2021).
- Gensim: Es la librería más rápida para el entrenamiento de vectores, puede procesar corpora arbitrariamente grandes, utilizando algoritmos de flujo de datos (Randim, 2009).
- TWilBert: Es un modelo de lenguaje neural bidireccional capaz de calcular representaciones de palabras bidireccionales gracias al uso de un modelo de lenguaje enmascarado (MLM) como objetivo de pre-entrenamiento (Moreno et al, 2021).

Para estos artículos de noticias se realizaron dos modificaciones en el procesamiento de la información:

- Cientos de palabras se fusionaron en forma de camel case (palabra1Palabra2); por lo tanto, antes de realizar el preprocesamiento, se separó cualquier palabra en mayúscula (ejemplo: "politicalParty" por "political Party").



- Se añadió una nueva variante de procesamiento con el fin de fusionar los procesamientos En bruto y Lematizado. A partir de ahí, los resultados fueron 5 modelos entrenados con Fasttext, 5 modelos entrenados con Gensim y 1 modelo entrenado con TWilBert (no se consideraron más experimentos por su bajo rendimiento).

Un paso importante en la generación de los EP fue averiguar la capacidad de representación semántica. Para ello, se ejecutó una validación con métodos intrínsecos sobre los embebimientos (Swysen, 2020).

4. Resultados

En este apartado se exponen los resultados obtenidos para cada uno de los métodos propuestos:

Lexicones

La Tabla 3 presenta los resultados para la clasificación de polaridad del conjunto de datos de entrenamiento utilizando la colección descrita en la sección *Lexicones*. Allí, iSOL, ElhPolar y Final Lexicon obtienen los mejores resultados en comparación con los lexicones de sentimiento en español. En general, la mayoría de los resultados están por encima de la media, excepto en SSL, cuya exactitud no superó el 33 %. Se destaca que "Lexicones final" es aquel que obtiene los mejores resultados con una exactitud total del 56 %.

Tabla 3. Resultados basados en el enfoque de lexicones

		ElhPo-				
		iSOL	lar	SSL	MS	FL
Precisión	Pos	82%	70%	76%	79%	81%
	Neg	65%	65%	64%	57%	57%
	Neu	51%	51%	50%	45%	45%
Exhaustividad	Pos	43%	35%	27%	34%	38%
	Neg	61%	77%	41%	78%	79%
	Neu	32%	32%	38%	40%	45%
Puntuación F1	Pos	57%	50%	40%	48%	52%
	Neg	63%	71%	50%	66%	66%
	Neu	45%	40%	47%	45%	43%
Exactitud		51%	53%	33%	53%	56%

Machine Learning

Los modelos con mayor exactitud son los clasificadores "MultinomialNB" y el "BernoulliNB", ambos con una exactitud del 60%; sin embargo, al comparar la puntuación F1, podemos concluir que Bernoulli NB es el modelo que tiene mejores resultados, a pesar de tener la misma precisión del Multinomial NB. Es importante mencionar que sólo se muestran los mejores resultados del tipo de datos para simplificar. En este caso, los mejores resultados están en preprocesamiento "En bruto".



Tabla 5. Resultados basados en el enfoque de Machine Learning

		Mult NB	SGD	Bernoulli	KNN
Precisión	Pos	58%	57%	58%	50%
	Neg	63%	63%	65%	68%
	Neu	100%	69%	62%	0%
Exhaustividad	Pos	90%	90%	89%	98%
	Neg	48%	44%	46%	20%
	Neu	0%	6%	15%	0%
Puntuación F1	Pos	71%	70%	70%	66%
	Neg	55%	52%	54%	31%
	Neu	1%	11%	24%	0%
Exactitud		60%	57%	60%	52%

Deep learning con embebimiento de palabras

Aquí se presenta Fasttext y Gensim con un embebimiento genérico pre entrenado (Wikipedia). El mejor resultado se consiguió con Gensim, que obtuvo una exactitud del 66,64%. Ahora bien, si comparamos cada resultado de Gensim contra los de Wikipedia podemos ver que esta última sólo es superior en un caso: En bruto. Por otro lado, los resultados de Fasttext no estaban lejos de los otros dos, siendo inferiores a lo que se decidió, como paso final, utilizar los embebimientos de Gensim con los datos procesados en formato "En bruto" y "En bruto lematizado". La cobertura se refiere a la proporción de palabras que están en el embebimiento y que al mismo tiempo están en los datos de entrenamiento.

Tabla 6. Resultados basados en el enfoque de Deep Learning

		Cobertura	Exac, Training	Exac, Testing
Gensim Word	Lematizado	54,08%	98,76%	64,92%
	Lematizado Normalizado	56,06%	98,26%	64,28%
	En bruto	61,99%	98,84%	66,64%
	En bruto lematizado	61,01%	98,94%	66,35%
	Normalizado	64,15%	98,14%	65,14%
Wikipedia	Lematizado	88,53%	94,92%	63,85%
	Lematizado Normalizado	86,88%	93,52%	63,56%
	En bruto	92,10%	95,56%	64,78%
	En bruto lematizado	88,53%	94,93%	62,84%
	Normalizado	90,06%	96,26%	65,78%
fastText	Lematizado	54,27%	99,13%	63,21%
	Lematizado Normalizado	56,28%	97,92%	62,20%
	En bruto	41,95%	97,95%	64,07%
	En bruto lematizado	52,66%	98,71%	60,63%
	Normalizado	64,15%	99,23%	65,00%

5. Conclusiones y trabajo futuro

Después de comparar los resultados obtenidos con los enfoques propuestos, se puede concluir que el mejor modelo es el enfoque basado en DL. Aunque la construcción de este modelo es más elaborada que la de los otros enfoques, los resultados son significativamente mejores que el enfoque basado en los lexicones y ligeramente mejores que el enfoque ML. Además, los resultados con las



embebimientos creadas fueron muy similares a los obtenidos con las preformadas en Wikipedia, a pesar de tener un 30% menos de cobertura. Esto significa que, al contarse con más datos para aumentar la cobertura, se puede mejorar la exactitud.

Como trabajo futuro, se propone el entrenamiento de un modelo DL con EP individuales específicas del español de cada país, esto permitirá que el modelo sea aún más preciso, teniendo en cuenta que hay palabras homónimas, pero tienen un significado diferente dependiendo del contexto en el que se escribieron, además de la forma de identificar los rasgos importantes, que en Puertas et al, (2021) mencionan como sintáctico, el simbólico y el complementario.

6. Referencias

Artículos de revistas

- Chandra, R., & Krishna, A. (2021). COVID-19 sentiment analysis via deep learning during the rise of novel cases. *PLOS ONE*, 16(8 August).
- Chiruzzo, L., Etcheverry, M., & Rosá, A. (2020). Sentiment analysis in Spanish tweets: Some experiments with focus on neutral tweets. *Procesamiento de Lenguaje Natural*, 64, 109–116
- Kamyab, M., Liu, G., & Adjeisah, M. (2021). Attention-Based CNN and Bi-LSTM Model Based on TF-IDF and GloVe Word Embedding for Sentiment Analysis. *Applied Sciences (Switzerland)*, 11(23).
- Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40(18), 7250–7257. <https://doi.org/10.1016/j.eswa.2013.06.076>
- Moreno-Sandoval, L. G., Pomares-Quimbaya, A., & Alvarado-Valencia, J. A. (2021). Celebrity profiling through linguistic analysis of digital social networks. *Computational Social Networks*, 8(1). <https://doi.org/10.1186/s40649-021-00097-w>
- Osorio Angel, S., Peña Pérez Negrón, A., & Espinoza-Valdez, A. (2021). Systematic literature review of sentiment analysis in the Spanish language. *Data Technologies and Applications*, 55(4), 461–479. <https://doi.org/10.1108/DTA-09-2020-0200>
- Puertas, E., Moreno-Sandoval, L. G., Redondo, J., Alvarado-Valencia, J. A., & Pomares-Quimbaya, A. (2021). Detection of Sociolinguistic Features in Digital Social Networks for the Detection of Communities. *Cognitive Computation*, 13(2), 518–537.
- Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36, 10–25.
- van Paridon, J., & Thompson, B. (2021). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, 53(2), 629–655.

Libros

- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python O'Reilly Media*.
- Nankani, H., Dutta, H., Shrivastava, H., Rama Krishna, P. V. N. S., Mahata, D., & Shah, R. R. (2020). *Multilingual Sentiment Analysis* (pp. 193–236).

Memorias de congresos

- Graves, A., & Schmidhuber, J. (2005). Framewise Phoneme Classification with Bidirectional LSTM Networks. *Proceedings of International Joint Conference on Neural Networks, Montreal*, 2047–2052.



- Harjule, P., Gurjar, A., Seth, H., & Thakur, P. (2020). Text Classification on Twitter Data. *Proceedings of 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things, ICETCE 2020*, 160–164.
- Moreno-Sandoval, L. G., Beltrán-Herrera, P., Vargas-Cruz, J. A., Sánchez-Barriga, C., Pomares-Quimbaya, A., Alvarado-Valencia, J. A., & García-Díaz, J. C. (2017). CSL: A Combined Spanish lexicon: Resource for polarity classification and sentiment analysis. *ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems*, 1, 288–295.
- Ochoa-Luna, J., & Ari, D. (2018). Deep neural network approaches for Spanish sentiment analysis of short texts. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11238 LNAI, 430–441.
- Tan, Y. (2018). An Improved KNN Text Classification Algorithm Based on K-Medoids and Rough Set. *Proceedings - 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2018*, 1, 109–113.
- Utiñaj, I., Morillo, P., & Huang, D. V. (2020, December 24). Sentiment Analysis Tool for Spanish Tweets in the Ecuadorian Context. *ACM International Conference Proceeding Series*.
- Villena, J., Daedalus, R., Lana-Serrano, S., Martínez-Cámara, E., & Carlos González-Cristóbal, J. (2013). TASS-Workshop on Sentiment Analysis at SEPLN TASS-Taller de Análisis de Sentimientos en la SEPLN. *Procesamiento Del Lenguaje Natural*, 50, 37–44.

Fuentes electrónicas

- Data Science Lab. (n.d.). *Multilingualsentiment*. Sites.Google.Com. Retrieved May 18, 2022, from <https://sites.google.com/site/datascienceslab/projects/multilingualsentiment>
- Facebook Inc. (2022). *fastText: Word representations*. Fasttext.Cc/. <https://fasttext.cc/docs/en/unsupervised-tutorial.html>
- Pedregosa, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot and Édouardand, M., Duchesnay, and Édouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12). <http://scikit-learn.sourceforge.net>.
- Perez, V., Banea, C., & Mihalcea, R. (n.d.). *Sentiment Lexicons in Spanish*. Web.Eecs.Umich.Edu. Retrieved May 18, 2022, from http://web.eecs.umich.edu/~mihalcea/downloads.html#SPANISH_SENT_LEXICONS
- Randim řehůřek. (2009). *GENSIM: topic modelling for humans*. Radimrehurek.Com. https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html#sphx-glr-auto-examples-tutorials-run-doc2vec-lee-py
- Roales González, N. (2014), *DETECCIÓN DE TENDENCIAS EN TWITTER UTILIZANDO MINERÍA DE DATOS ADAPTATIVA*, Universidad Autónoma de Madrid. https://repositorio.uam.es/bitstream/handle/10486/662510/roales_gonzalez_natalia_tfg.pdf?sequence=1
- Secretaria de Estado de Investigación Desarrollo e Innovación. (2011). *iSOL*. Timm.Ujaen.Es. <http://timm.ujaen.es/recursos/isol/>
- Swysen, T. (2020). Swysen2020. Universidad de Chile.

Sobre los autores

- **Luis G. Moreno Sandoval:** Ingeniería Sistemas, Máster Dirección comercial y Marketing Digital, MBA, Maestría en Ciencias de la Información y Comunicaciones, Candidato



Doctoral en Ingeniería de Pontificia Universidad Javeriana. Profesor cátedra. morenoluis@javeriana.edu.co

- **Alexandra Pomares Quimbaya:** Ingeniería de Sistemas, Maestría en Ingeniería de Sistemas y Computación, Doctorado en Informática en Universidad Grenoble, Doctorado en Ingeniería de Universidad de los Andes. Profesor titular. pomares@javeriana.edu.co
- **Camilo Eduardo Cruz Gutiérrez:** Ingeniería de Sistemas. cruzcamilo@javeriana.edu.co
- **Johan Ferney García Pachón:** Ingeniería de Sistemas. johan-garcia@javeriana.edu.co
- **David Felipe Vanegas Ramírez:** Ingeniería de Sistemas. david-vanegas@javeriana.edu.co

Los puntos de vista expresados en este artículo no reflejan necesariamente la opinión de la Asociación Colombiana de Facultades de Ingeniería.

Copyright © 2022 Asociación Colombiana de Facultades de Ingeniería (ACOFI)

