Modelado e implementación de algoritmos inteligentes de análisis de opinión

Resumen de la Tesis presentada para obtener el grado de Doctor en Ciencias Informáticas

Posgrado, Facultad de Informática, Universidad Nacional de La Plata (UNLP)

Fecha de defensa: 27/09/2023

Autor: Juan Pablo Tessore juanpablo.tessore@itt.unnoba.edu.ar [1 2]

Directora: Sandra Baldassarri sandra@unizar.es [3]

Codirector: Hugo Ramón hugo.ramon@itt.unnoba.edu.ar [1]

[1] Instituto de Investigación y Transferencia en Tecnología (ITT) – Centro Asociado a la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC) / Escuela de Tecnología / Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA)

[2] Comisión Nacional de Investigaciones Científicas y Técnicas (CONICET)

[3] Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza (UNIZAR)

Resumen

Numerosos estudios que abordan el Análisis de Sentimientos, mediante clasificadores basados en aprendizaje automático, se enfrentan a la necesidad de etiquetar manualmente los datos de entrenamiento o utilizar conjuntos de datos públicos pre etiquetados. Estos enfoques presentan desafíos significativos, como la laboriosidad del etiquetado manual, la ambigüedad en la clasificación de emociones y la falta de recursos en idiomas distintos al inglés.

Con el propósito de agilizar el desarrollo de clasificadores de emociones en textos en español, basados en aprendizaje automático supervisado, esta tesis propone y ejecuta una metodología para la captura, etiquetado y validación del contenido, con el fin de acelerar el proceso de construcción del conjunto de datos. Además, se desarrollan clasificadores basados en aprendizaje automático utilizando los datos recopilados y se compara su rendimiento con el de clasificadores entrenados con datos etiquetados manualmente. Por último, se evalúa el impacto de la información contextual en el rendimiento de los clasificadores construidos.

Palabras clave

Análisis de Sentimientos, Procesamiento del Lenguaje Natural, Supervisión Distante, Aprendizaje Automático, Información Contextual.

Motivación

La interacción humana en las redes sociales es ahora un aspecto fundamental de la vida cotidiana, facilitando la expresión de opiniones y emociones sobre una amplia gama de temas. Este flujo constante de interacciones ofrece una valiosa fuente de información para comprender eventos, productos y servicios. El Análisis de Sentimientos y la Computación Afectiva han surgido como áreas de investigación clave para abordar esta avalancha de datos sociales. Estas disciplinas se centran en comprender y procesar las emociones humanas, aprovechando la interacción humano-computadora y las señales multimodales [1-4].

Sin embargo, a pesar del crecimiento en esta área, la mayoría de las investigaciones se han centrado en el inglés, dejando un vacío significativo para otros idiomas, como el español. La falta de recursos etiquetados en español ha impulsado la necesidad de métodos más eficientes para construir conjuntos de datos de Análisis de Sentimientos. La supervisión distante (DS *distant supervisión* por sus siglas en inglés) ha surgido como una alternativa prometedora, permitiendo la recopilación automática de datos etiquetados [5]. Sin embargo, la calidad de estos conjuntos de datos es un desafío, especialmente dada la naturaleza subjetiva de las emociones.

En este contexto, este trabajo de tesis propone una metodología para la construcción y validación de recursos de Análisis de Sentimientos en español, centrándose en la detección de emociones básicas [6]. Se reconoce la importancia de la validación de la calidad de los conjuntos de datos recopilados mediante DS, especialmente en lo que respecta al nivel de consenso entre las etiquetas asignadas y el rendimiento de los clasificadores basados en aprendizaje automático (ML *machine learning* por sus siglas en inglés).

Además, se destaca la necesidad de fomentar la detección de emociones en español, un área que ha recibido menos atención en comparación con la clasificación de polaridad. La creación de conjuntos de datos etiquetados específicamente en español es un primer paso crucial para avanzar en esta dirección [7]. Finalmente, se plantea la cuestión de si la información contextual (IC) asociada a los datos etiquetados puede mejorar el rendimiento de los clasificadores, lo que sugiere una posible área de investigación futura [8].

Aportes y contribuciones

Esta tesis constituye una contribución significativa al campo del Análisis de Sentimientos al presentar una metodología detallada para la creación de recursos específicos de emoción básica, centrándose particularmente en el idioma español, aunque su aplicabilidad se extiende a otros idiomas. Enfatiza la importancia de abordar la tarea de detección de emociones básicas debido a las carencias existentes en este ámbito para el español.

La metodología propuesta se fundamenta en el uso de técnicas de DS, lo que permite una recopilación eficiente de datos etiquetados, minimizando la intervención humana y posibilitando la obtención de conjuntos de datos más grandes en comparación con los obtenidos mediante etiquetado manual. Sin embargo, para garantizar la calidad de estos conjuntos de datos, se incorpora una fase de validación que incluye un muestreo del conjunto de datos, el etiquetado manual de la muestra y la

comparación de las etiquetas asignadas manualmente con las originales, utilizando métricas de consenso, como el Kappa de Fleiss [9].

Es relevante destacar que la metodología propuesta no solo se enfoca en la creación y validación de los conjuntos de datos, sino que también aborda la implementación de clasificadores de emoción básica basados en ML. Esta es una contribución significativa, ya que la mayoría de los trabajos existentes carecen de una metodología clara para este proceso en el contexto del español.

Descripción general del proceso

En la Figura 1 se presentan las distintas etapas del proceso para el Análisis de Sentimientos desarrollado a lo largo de esta tesis.

Dicho proceso se basa en la arquitectura general para sistemas de minería de texto [10], e incluye las etapas recopilación, preprocesado y validación de conjuntos de datos junto con la selección de formatos de representación y algoritmos de clasificación.

En la primera etapa del proceso, se realiza la recopilación de los datos, seleccionando

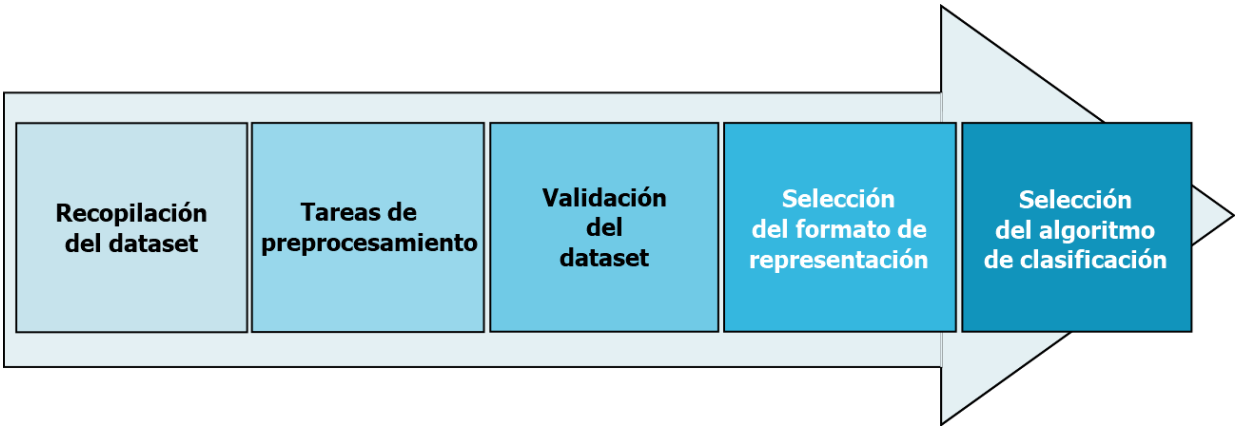

Figura 1: Descripción general del proceso para el Análisis de Sentimientos desarrollado.

El resultado final de la aplicación de esta metodología es la generación de un conjunto de datos etiquetado y validado, que puede ser utilizado como punto de partida para futuras investigaciones en el campo del Análisis de Sentimientos en español. Además, al incorporar IC mediante la DS, se observa una mejora en el desempeño de los clasificadores implementados, lo que resalta la eficacia de esta metodología en la generación de recursos de alta calidad para el análisis de emociones en textos en español y potencialmente en otros idiomas.

una fuente que contenga mayoritariamente textos en español en la variante de Argentina, y que permita asociarlos de manera sencilla a un marcador asimilable a una emoción básica. De esta manera, se busca construir un conjunto de datos con un tamaño de muestras considerable (en la escala de los cientos de miles), utilizando el enfoque conocido como DS. Para la selección de la fuente de datos, también se considera que la misma incluya IC asociada al contenido, de manera de utilizarla posteriormente para intentar mejorar el desempeño del clasificador basado en ML a construir.

A continuación, se realizan las distintas tareas de preprocesamiento sobre el conjunto de datos: tokenización, filtrado de muestras por longitud, filtrado y descarte de muestras en otro idioma y filtrado de comentarios pertenecientes a trolls. Una vez realizadas dichas tareas, se compara el desempeño de clasificadores basados en ML utilizando el conjunto de datos original (esto es, sin tareas de preprocesamiento) y el revisado, de manera de establecer la efectividad de las tareas desarrolladas para los datos recopilados. Este paso resulta necesario debido a que no necesariamente todas las tareas de procesamiento resultan efectivas independientemente del conjunto de datos.

Posteriormente, debido a que el conjunto de datos fue construido mediante un enfoque de DS, se encara una validación sobre las etiquetas asignadas al contenido. En esta etapa, no resulta posible validar la totalidad de las muestras dado que el tiempo que insumiría tal proceso diluiría las ventajas de la DS. Por el contrario, se realiza una validación estadística tomando un conjunto reducido de muestras, las cuales se clasifican de manera manual por un grupo de psicólogos, para luego calcular la métrica Kappa de Fleiss para las etiquetas asignadas por los psicólogos entre sí y entre las anteriores respecto a la etiqueta original de contenido.

Por último, se selecciona un formato de representación de textos, que sirva como entrada para el clasificador basado en ML para el cual también debe seleccionarse un algoritmo. En ambos casos, se pondera un formato y algoritmo que sean susceptibles de ser mejorados con el uso de IC, pero a su vez que sea posible comparar la mejora en el desempeño con otros estudios presentes en la bibliografía.

Construcción y validación del conjunto de datos

Se optó por utilizar la red social Facebook como plataforma principal para construir un conjunto de datos de emoción básica, con el objetivo de que las reacciones de Facebook reflejen las emociones expresadas por los propios usuarios. Dada la enorme cantidad de usuarios activos en Facebook en ese momento, se consideró una fuente de datos ideal para este propósito. Además, se eligieron cuidadosamente 13 portales de noticias populares en Argentina, como Clarín, La Nación y Página12, para recopilar una variedad de contenido. Se recolectaron títulos de noticias, descripciones, comentarios y reacciones de los usuarios para cada publicación. La asociación entre los comentarios y las reacciones permitió vincular la emoción expresada por el usuario con la noticia en cuestión. Esta asociación se basó en la correspondencia entre ciertas reacciones (como "ANGRY" y "SAD") y las emociones básicas propuestas por Ekman [11].

Se recopilaron un total de 20,996,169 comentarios en la fase inicial, pero solo se seleccionaron aquellos con reacciones asociadas para su utilización como etiquetas. De las reacciones disponibles, "LOVE", "HAHA", "ANGRY" y "SAD" se consideraron para el conjunto de datos, mientras que "WOW" se excluyó debido al debate sobre su inclusión como emoción básica. Los comentarios asociados con la reacción "LIKE" también se excluyeron, ya que su significado es ambiguo. En caso de múltiples comentarios del mismo usuario en una publicación, se seleccionó el comentario más antiguo para asociarlo con la reacción, considerando que reflejaba mejor su estado emocional al leer la noticia. Tras estas

consideraciones, el número de comentarios etiquetados se redujo a 1,716,413.

Posteriormente, se aplicaron técnicas de preprocesamiento, como la tokenización y filtrado de comentarios en otros idiomas o de longitud insuficiente. La tokenización se realizó utilizando la clase TweetTokenizer de NLTK para eliminar tokens no útiles y stopwords en español. Los comentarios con menos de tres tokens válidos se filtraron, reduciendo el número a 1,261,783. Se validó el idioma de los comentarios, eliminando aquellos que no estaban en español, lo que redujo el total a 1,035,045. Además, se aplicó un proceso de validación cruzada utilizando Google Translate para confirmar

los resultados obtenidos. Se eliminaron comentarios potencialmente de trolls, identificados como aquellos que se repetían frecuentemente en varias publicaciones. Después de este filtro, quedaron 1,020,557 comentarios útiles para el análisis.

Después de seleccionar y filtrar el contenido relevante del conjunto de datos, es esencial abordar la etapa de validación para asegurar la fiabilidad de las etiquetas asignadas. Para ello, se empleó la métrica Kappa de Fleiss, comúnmente utilizada en conjuntos de datos construidos manualmente, para medir el nivel de consenso entre las etiquetas asignadas. Se considera un nivel de consenso moderado en la escala de Kappa de Fleiss como criterio para determinar la idoneidad de los datos para la construcción de clasificadores basados en ML.

Dado que la validación manual completa de los datos sería impracticable y contraproducente, se optó por una muestra representativa. Este enfoque se ha utilizado en investigaciones previas, donde se tomaron muestras que representaban menos del 0.02% del total del conjunto de datos para su validación. Siguiendo este principio, se estimó el tamaño de la muestra necesario y se llevó a cabo un proceso de re etiquetado por especialistas en psicología. Finalmente, se compararon las etiquetas originales con las proporcionadas por los especialistas para calcular los valores de Kappa de Fleiss y se determinó si se cumple con el requisito de un consenso moderado.

Métrica	Acuerdo
Fleiss Kappa global	0,4426
Fleiss Kappa ANGRY vs todos	0,4071
Fleiss Kappa HAHA vs todos	0,4415
Fleiss Kappa LOVE vs todos	0,5452
Fleiss Kappa SAD vs todos	0,4081

Tabla 1:Acuerdo entre los etiquetadores humanos y la etiqueta original del conjunto de datos.

La muestra resultante fue de 247 comentarios, divididos en diez conjuntos de 24 o 25 cuádruplas. Cada conjunto se utilizó para crear un formulario en Google con una tarea de clasificación por cuádrupla. Este enfoque permitió una validación estadística efectiva del conjunto de datos sin la necesidad de un re etiquetado manual completo. Para la clasificación manual de la muestra se pidió la colaboración de 25 psicólogos.

Métrica	Acuerdo
Fleiss Kappa global	0,4911
Fleiss Kappa ANGRY vs todos	0,4933
Fleiss Kappa HAHA vs todos	0,4989
Fleiss Kappa LOVE vs todos	0,5332
Fleiss Kappa SAD vs todos	0,4240

Tabla 2: Acuerdo entre etiquetadores humanos.

El resultado del proceso de re etiquetado puede verse en la Tabla 1 y Tabla 2. Aquí se destaca que todas las métricas se encuentran en el rango moderado de la escala de Kappa de Fleiss. Dicho rango fue considerado apto

por otros investigadores para entrenar clasificadores basados en ML [12 – 13]

Construcción de clasificadores y utilización de información contextual

En la presente sección, se desarrollan las dos últimas etapas del proceso que se muestra en la Figura 1. Estas son en primer lugar la selección de un formato de representación para el texto y, en segundo lugar, la selección de un algoritmo de ML.

Por otro lado, entre los algoritmos de clasificación más comunes, las redes neuronales recurrentes (RNN) y sus variantes, como las LSTM, son especialmente útiles para el procesamiento de datos secuenciales en el análisis de sentimientos. Aunque los conjuntos de datos pueden diferir en su naturaleza, como los diálogos textuales versus los comentarios en redes sociales, se considera que estas diferencias no son críticas, ya que ambos son formas de comunicación o diálogos.

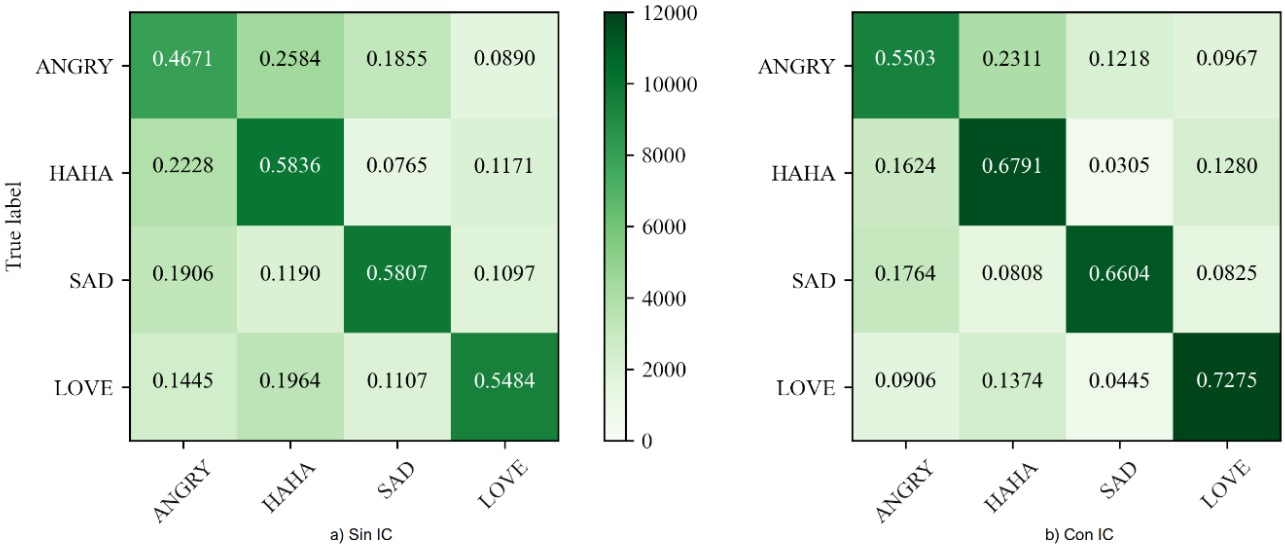

Figura 2: Matrices de confusión para los resultados de pruebas.

El formato de representación en el análisis de sentimientos ha evolucionado, mostrando limitaciones en enfoques tradicionales como las bolsas de palabras o caracteres, especialmente para textos cortos como tweets. Por ello, muchos estudios recientes han adoptado embeddings neuronales como Word2Vec, FastText y GloVe, que capturan el contexto de las palabras en el mismo embedding. Aunque inicialmente estos recursos se desarrollaron para el inglés, se han creado recursos similares para el español.

La elección de utilizar embeddings neuronales, especialmente GloVe, y redes LSTM para los experimentos se basa en su actualidad y en la necesidad de comparar el desempeño del clasificador con una línea base. Se tomó como referencia un trabajo relevante que utiliza una arquitectura similar y aborda objetivos similares [13]. Sin embargo, el presente trabajo de tesis se enfoca en medir el impacto de la DS y la IC en el rendimiento del clasificador en lugar de construir el clasificador óptimo posible.

Se llevaron a cabo dos experimentos, el primero para medir el impacto de la IC en una

tarea de clasificación de emoción básica y el segundo para establecer cómo se comporta un conjunto de datos recopilado mediante DS en comparación con una contraparte clasificada manualmente. El formato de embedding utilizado fue GloVe, y el algoritmo de clasificación seleccionado fue LSTM.

Los vectores de GloVe (proporcionados por el proyecto *Spanish Billion Words Corpus*) se usaron para construir una capa de embeddings no entrenable. El tamaño del vocabulario se limitó a 20.000 palabras. Luego, la capa anterior se conectó con una capa LSTM de 128 dimensiones con un dropout de 0,2 para evitar el overfitting. Por último, se añadió una capa densa con la función sigmoide como activación.

En la Figura 2 pueden verse los resultados del uso de IC en el proceso de clasificación. La matriz de confusión correspondiente al clasificador entrenado con IC muestra un mejor desempeño con respecto al entrenado sin ella.

Por otra parte, en la Tabla 3, puede compararse el desempeño del clasificador entrenado con los datos recopilados mediante DS en el presente trabajo de tesis, contra un clasificador entrenado con los datos de la competencia SemEval-2019-Task 3, el cual fue etiquetado de manera manual. El clasificador entrenado con los datos de esta tesis logra un desempeño comparable al entrenado con datos de etiquetado manual en términos de F1, pero además es más sensible al uso de IC.

	SemEval-2019-Task 3*			**Este trabajo**		
	Sin Contexto	**Con Contexto**	**Mejora (%)**	**Sin Contexto**	**Con Contexto**	**Mejora (%)**
Accuracy	0,8403	0,8448	0,5355	0,4742	0,5444	14,8039
Precision	0,4755	0,4719	-0,7571	0,4783	0,5570	16,4541
Recall	0,6689	0,7215	7,8637	0,4740	0,5437	14,7046
F1	0,5559	0,5706	2,6444	0,4762	0,5503	15,5607

Tabla 3: Comparativa de desempeño entre clasificadores entrenado mediante distintos conjuntos de datos.

Conclusiones

Las principales contribuciones de este trabajo de tesis incluyen el desarrollo de una metodología para el Análisis de Sentimientos y la creación de conjuntos de datos de emoción básica, así como su aplicación en el entrenamiento de clasificadores basados en ML. Se recopiló dataset de emoción básica en español, superando en tamaño a los existentes en la literatura.

La metodología permitió obtener conclusiones específicas en cada etapa del proceso. Se demostró que las tareas de limpieza y preprocesamiento no garantizan una mejora significativa en el desempeño de los clasificadores, aunque son indispensables para la validación del conjunto de datos.

Por otro lado, la etapa de validación incluyó un muestreo del conjunto de datos y re etiquetado por expertos, revelando que es posible lograr un nivel de consenso moderado, medida por la Kappa de Fleiss, aceptable para la construcción de clasificadores.

La construcción de los clasificadores confirmó su sensibilidad al tamaño del conjunto de datos, mostrando mejoras significativas al utilizar IC. Esto destaca la importancia de prestar atención a la captura de IC en la metodología. Además, se demostró que los clasificadores entrenados con conjuntos de datos construidos mediante DS pueden alcanzar rendimientos comparables a los de conjuntos etiquetados manualmente, como se observa los experimentos detallados en la sección anterior.

Líneas de investigación futuras

Los pasos futuros de esta investigación se centran en diversas áreas clave. En primer lugar, es necesario probar la metodología propuesta para el Análisis de Sentimientos en una variedad de conjuntos de datos, tipos de datos de entrada, idiomas y algoritmos de clasificación. Esto permitiría verificar la eficacia de las tareas de preprocesamiento y determinar si las mejoras observadas son consistentes en diferentes contextos.

Además, se sugiere explorar el uso de distintos tipos de etiquetas para clasificar automáticamente el contenido, así como también considerar la inclusión de IC para mejorar el rendimiento de los clasificadores. Se plantea la necesidad de repetir los experimentos para distintos idiomas.

Otro aspecto importante es la incorporación de formatos de entrada multimodales, como audio, video e imágenes, lo que requeriría adaptaciones en todas las etapas del proceso, desde el preprocesamiento hasta la construcción de clasificadores.

En cuanto al proceso de validación de etiquetas, se sugiere explorar diversas variaciones en los cuestionarios para medir su impacto en los resultados del consenso, como la longitud y el método de re etiquetado.

Para optimizar el desempeño del clasificador, se propone experimentar con diferentes configuraciones, como las utilizadas en competencias relevantes y explorar el uso de información semántica y representaciones novedosas, como BERT.

Es importante reconocer que este trabajo tiene limitaciones, especialmente en relación con el uso de IC en conjuntos de datos de Análisis de Sentimientos. Se necesitan más pruebas para demostrar la consistencia de esta mejora y para optimizar aún más la metodología propuesta.

Referencias

1. Cambria, E. (2016). Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems*, *31*(2), 102–107. https://doi.org/10.1109/MIS.2016.31
2. Picard, R. (1997). Affective Computing. MIT Press.
3. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies (Vol. 5). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-02145-9
4. Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18–38. https://doi.org/10.1016/j.ins.2015.03.040
5. Roth, B., Barth, T., Wiegand, M., & Klakow, D. (2013). A survey of noise reduction methods for distant supervision. In AKBC 2013 - Proceedings of the 2013 Workshop on Automated Knowledge Base

Construction, Co-located with CIKM 2013 (pp. 73–77). San Francisco, California: Association for Computing Machinery. https://doi.org/10.1145/2509558.2509571

6. García-Vega, M., Díaz-Galiano, M. C., García-Cumbreras, M., Del Arco, F. M. P., Montejo-Ráez, A., Jiménez-Zafra, S. M., … Moctezuma, D. (2020). Overview of TASS 2020: Introducing Emotion Detection. In CEUR Workshop Proceedings (Vol. 2664, pp. 163–170).
7. 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
8. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems (Vol. 89). https://doi.org/10.1016/j.knosys.2015.06.015
9. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378–382. https://doi.org/https://doi.org/10.1037/h0031619
10. Feldman, R., & Sanger, J. (2006). The Text Mining Handbook. The Text Mining Handbook. Cambridge University Press. https://doi.org/10.1017/CBO9780511546914
11. Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124–129. https://doi.org/10.1037/h0030377
12. Gambino, O. J., & Calvo, H. (2019). Predicting emotional reactions to news articles in social networks. Computer Speech & Language, 58, 280–303. https://doi.org/10.1016/j.csl.2019.03.004
13. Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019). SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 39–48). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2005