XIV Jornadas de Sociología. Carrera de Sociología, Facultad de Ciencias Sociales, UBA.

Eje 1. Filosofía, Teoría, Epistemología, Metodología

Mesa 254. Ciencias sociales computacionales y big data

Hacia un análisis de la polaridad del Big Data

Christian Ratovicius

(Universidad Abierta Interamericana)

cratovicius@yahoo.com

Resumen: Investigaciones sociales en torno a desarrollos tecnológicos recientes, tales como el “big data” o la inteligencia artificial, señalan una cierta ambivalencia o bipolaridad en los distintos ámbitos de comunicaciones donde los refieren, o incluso en el sentido común de los distintos grupos sociales. Así, por ejemplo, hay una retórica favorable a estas tecnologías que las identifica con algunas oportunidades para distintos sectores comerciales, tales como la industria 4.0, o con los beneficios que podría aportar en el campo médico; a la vez que un discurso crítico señala los riesgos que con llevan en materia de avance sobre la privacidad, manipulación mediática y política, o en el origen de nuevas desigualdades o situaciones de injusticia. En este trabajo proponemos un análisis de polaridad sobre un corpus de oraciones que incluyen el término “big data”, construido a partir de noticias recogidas de periódicos online argentinos. Particularmente, proponemos dos abordajes que combinados permitirán un análisis más robusto: Análisis por medio del uso de lexicones y diccionario, otro por medio del uso de clasificadores. Este trabajo nos va a permitir clasificar las oraciones para profundizar el entendimiento de esta temática.

1. Introducción

El “big data” suele ser representado tanto como un fenómeno positivo tanto como negativo (Kitchin, 2014). Cuando se lo representa positivamente generalmente se lo vincula con posibles beneficios o promesas en los distintos sectores en los que se inserta, tales como una revolución productiva en la industria 4.0, el desarrollo de ciudades inteligentes y un mejor manejo de los problemas urbanos, entre otros. La narrativa positiva, sin embargo, convive con una que hace foco en sus riesgos, entre los que se destaca la explotación con fines comerciales o políticos de los datos personales. Por estas razones Paganoni (2019) sugiere que el big data se encuentra tensionado por

dos polos: el de los datos y la información, enfrentado al de los derechos y la privacidad. Buscando aportar a esta línea de estudio, en este trabajo reportamos los avances preliminares del desarrollo de una herramienta para identificar la polaridad de un enunciado sobre el big data.

La tarea propuesta cae entonces dentro del campo denominado análisis de sentimiento o minería de opinión, una rama del procesamiento de lenguaje natural (NLP) que estudia la identificación y extracción de la información subjetiva de un texto, con el fin de analizar opiniones, actitudes, valoraciones, emociones y sentimientos hacia diferentes entes y sus atributos expresados en un texto escrito (Liu, 2015).

Este tipo de análisis pueden ser entendidos como procesos de clasificación, de los cuales conviene distinguir dos: a nivel de documento u oración, y a nivel de los objetos u aspectos (Medhat, Hassan, & Korashy, 2014). Los primeros consideran al documento o a la oración como un todo con una sola unidad básica de información, o como si tuvieran sólo un tema; a nivel de entidades y aspectos la tarea es más compleja ya que se deben distinguir distintos objetos u aspectos de los objetos, antes de poder indagar cuál es la valoración de cada uno.

La necesidad de optar por un tipo de sistemas u otro se vincula a su vez al tipo de corpus con el que se trabaja. Consideremos, por ejemplo, un corpus construido por oraciones relativamente simples que claramente refieren al “big data”, tales como “big data es una gran herramienta para agregar valor a los servicios al cliente” o “big data significa muchos datos”. En este corpus el desafío se limita a la identificación de la polaridad. Otro tipo de corpus puede estar construido de manera tal que incluya enunciados más complejos, donde el “big data” no es el principal objeto de la oración, como “la reunión tratará sobre big data e inteligencia artificial” o “se espera un crecimiento dijo el responsable de big data”. En estos casos es necesario un análisis a nivel de entidades u aspectos.

En este trabajo nos limitamos al primer tipo de análisis (de polaridad), en el nivel de las oraciones. Por el momento, esto nos condiciona a trabajar con un corpus de oraciones de las cuales ya sabemos que tratan sobre el “big data”. Este filtro se podría automatizar, incluyendo tareas del otro tipo de análisis (sobre objetos u aspectos).

Para tal objetivo las técnicas se suelen clasificar en 2 enfoques:

- Las basadas lexicones o diccionarios utilizan un listado de palabras previamente calificadas con un valor numérico. Este diccionario suele ser cruzado con el texto a evaluar por medio de una función que dará como resultado un valor. De

acuerdo cómo se incluyan las reglas en dicha función, se considerará cada palabra contra el diccionario para devolver el valor, o algún tipo de combinación de ellas, con posibles ponderaciones. Una de las principales dificultades de este método es la confección del diccionario, ya que no hay diccionarios universales, y las valoraciones de las palabras pueden diferir según el contexto.

- Las basadas en aprendizaje automático. Estos se basan en la aplicación de algoritmos de machine learning para identificar patrones en casos ya clasificados (etiquetados) para desarrollar reglas que permitan clasificar nuevos casos. Dado que no se incluyen reglas explícitas, estos sistemas son muchos más flexibles. La principal dificultad de esta técnica es que requiere un buen conjunto de casos ya clasificados para poder generar reglas que se puedan generalizar.

Luego hay métodos híbridos que utilizan lo mejor de la combinación de ambos.

En este trabajo reportamos los primeros avances del desarrollo de una herramienta capaz de clasificar enunciados sobre big data, detallando su polaridad positiva o negativa, a través de los dos enfoques mencionados, dejando el enfoque mixto para futuros trabajos.

2. Metodología

En esta primera parte del trabajo del análisis de datos vamos a preparar los elementos para el análisis. Para todas ellas se usará lenguaje R (R Core Team, 2018).

2.1. Construcción de corpus

Las oraciones se extrajeron de un corpus de 2.026 noticias publicadas en periódicos digitales argentinos que incluían explícitamente el término “big data” en su contenido. La construcción de ese corpus se detalla en Becerra (2018). Aquí se extrajeron las 2785 oraciones donde explícitamente se incluía dicho término. Este corpus es uno entre varios que constituyen una base mayor de enunciados sobre “big data”, incluyendo sus expresiones en otros tipos de fuentes, como artículos científicos, documentos políticos, o apariciones en redes sociales.

Estas oraciones fueron clasificadas a mano de la siguiente manera:

Variable	Frequencia
negativo	280
neutral	1279
positivo	1226

Sin embargo, se debe advertir que la categoría de “neutrales” incluye oraciones que no refieren indistinta y exclusivamente al “big data”, siendo esta una tarea pendiente por fuera del alcance de este trabajo.

2.2. Recursos y diccionarios

Para las tareas necesarias en el primer enfoque se construyó un lexicón, combinando dos diccionarios:

En primer lugar, se trabajó con el Spanish Dictionary Affect Language (en adelante “SDAL”), desarrollado por Gravano & Dell’Amerlina Ríos (2014), que replica el modelo de Whissell (2009). Este es un lexicón formado por 2880 términos, clasificados manualmente en tres dimensiones afectivas, de las cuales aquí utilizamos el agrado, el cual se expresa en un rango de 1 (negativo) a 3 (positivo) Es importante señalar que este diccionario incluye términos lemantizados y transformados para incluir un sufijo que denote el tipo de palabra mencionada, a fin de evitar ambiguedades (e.g., “material_N” denota el sustantivo de “material”, mientras “material_A” el adjetivo).

En segundo lugar, se trabajó con el Lexicón de evocaciones a “big data” (en adelante “EVOC”), desarrollado por Becerra & López-Alurralde (2020) en el marco de una investigación sobre las representaciones sociales del “big data” con la técnica de la evocación libre de palabras, a la que se añadió la posibilidad de aclarar la valoración del término incluido. Este es un lexicón formado por 1516 términos. Este diccionario ya incluye los términos lemantizados. Las valoraciones están expresadas en escala de 0 (negativo) a 10 (positivo).

2.3. Preprocesamientos

Para poder utilizar los lexicones en conjunto se debió hacer un cierto preprocesamiento:

- En SDAL se eliminaron los sufijos, calculando la media entre los términos ambiguos.
- Tanto en SDAL como EVOC, se escalaron las valoraciones dentro del rango 0 (negativo) y 1 (positivo).

Entre los diccionarios se observó sólo un solapamiento de términos del 0%.

Para las tareas de clasificación con el segundo enfoque (aprendizaje automático), se anotó al texto con el parser sintáctico del paquete Udpipe Package for R, con su modelo de lenguaje para español (Straka & Strakova, 2017) y también se lo lemantizó, así como se convirtieron mayúsculas en minúsculas; luego, se filtraron sólo aquellas palabras que

eran adjetivos, verbos, sustantivos y adverbios, los primeros 3 por ser los más informativos y presentes en los corpus, y los últimos como modificadores del sentido de las oraciones;

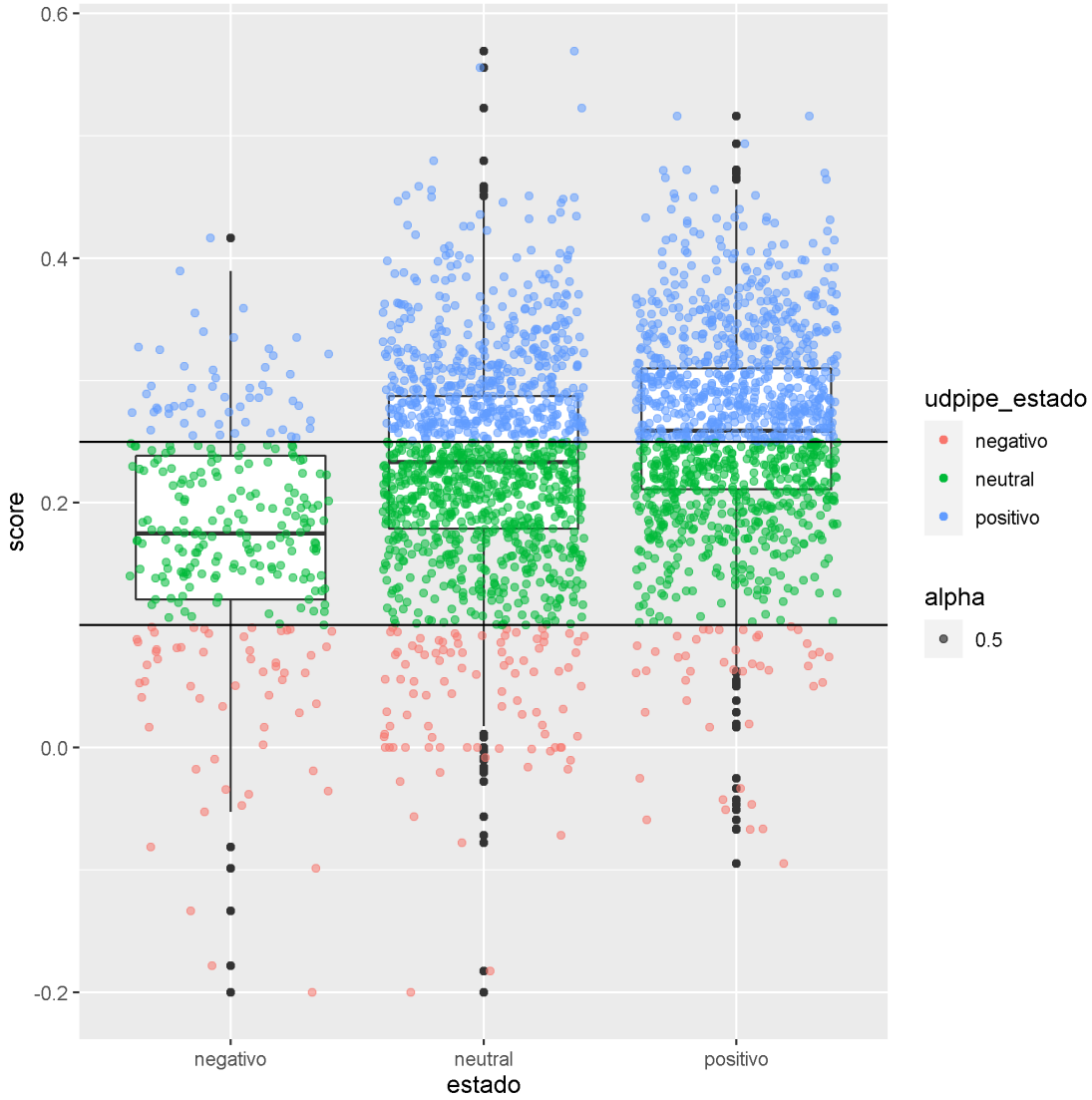
Al finalizar el preprocesamiento quedamos con corpus de 7920 lemas únicos, y diccionarios de 2880 (SDAL) y 1516 (EVOC).

3. Resultados

3.1. Sistema basado en diccionarios

Para la parte de análisis con léxicos utilizamos la función txt_sentiment del paquete UdPipe, que identifica palabras que tienen un significado positivo/negativo, con la adición de alguna lógica básica con respecto a las apariciones de amplificadores, desamplificadores y negadores en la proximidad de la palabra que tiene un significado positivo o negativo. Como esta función trabaja con palabras sueltas, en esta etapa vamos a trabajar sólo con adjetivos, verbos, sustantivos, adverbios y artículos.

En los estudios anteriores, de donde tomamos el corpus de oraciones y el léxico EVOC, ya se había advertido que las palabras en el campo semántico del “big data” reflejan su discurso promocional, y por ello una cierta tendencia a una valoración más positiva. Por este corrimiento, anotamos el estado positivo cuando el score es mayor a 0.25 y consideramos negativas cuando son menores a 0.1.



Por último, vemos como seria la matriz de confusión donde se puede obtener los resultados comparados como la función calificó a las oraciones versus cómo se analizaron manualmente.

Esta estrategia parece ser mejor clasificando positivos y neutrales que negativos.

	negativo	neutral	positivo
negativo	56	170	54
neutral	93	653	533
positivo	45	495	686

Se observa que clasifico en forma correcta el 21% de los negativos, el 51% de los neutrales y el 56% de los positivos.

3.2. Sistema basado en diccionarios

Como segunda técnica a utilizar la metodología de aprendizaje supervisado. Esto se refiere que partimos de datos ya etiquetados previamente, un “trainning dataset”, que busca inferir cómo clasificar a otros “test dataset”.

Nuestro plan de trabajo sigue las siguientes tareas, para las que seguimos a Hvitfeldt y Silge (2021).

- Pre-procesar texto.
- Vectorizar texto.
- Dividir el corpus para entrenamiento y testeo de este.
- Armar el modelo de clasificación.
- Evaluar el modelo
- Ajustar el modelo

Originalmente nuestros datos tienen las siguientes etiquetas:

Variable	Frequencia
negativo	280
neutral	1279
positivo	1226

Como podemos observar el grupo de oraciones a clasificar no es igual lo que reviste uno de los tantos problemas que pueden emerger en cualquier proyecto de ciencia de datos. En muchas situaciones, la cantidad y calidad de estos, no dependen de la responsabilidad del investigador, quien deberá pensar en estrategias para hacer su corpus más equitativo.

Luego, dividimos nuestros datos en training/test sets en una proporción de 5/6.

Para una mejor forma de trabajar dividimos los datos en grupos en tamaños aproximadamente iguales, esto se realiza en forma aleatoria esto se denominan FOLDS.

En la preparación, preprocesamos estos datos para prepararlos para el modelado, ya que tenemos datos de texto, y necesitamos construir características numéricas para el aprendizaje automático a partir de ese texto. El paquete de recetas, que forma parte de tidymodels, nos permite crear una especificación de los pasos de preprocesamiento que queremos realizar. Estas transformaciones se estiman (o se “entrenan”) en el conjunto de

entrenamiento para que puedan aplicarse de la misma manera en el conjunto de pruebas o en los nuevos datos en el momento de la predicción, sin que se filtren los datos. En primer lugar, convertimos el texto en palabras con step_tokenize(). Por defecto se utiliza tokenize_words(). Antes de calcular el tf-idf utilizamos step_tokenfilter() para mantener sólo los 3000 tokens más frecuentes, para evitar crear demasiadas variables en nuestro primer modelo. Para terminar, utilizamos step_tfidf() para calcular tf-idf.

En lo que sigue dividimos el modelo en 2: una se crea con la función step_upsample(), que crea una especificación de un paso de receta que replicará las filas de un conjunto de datos para igualar la aparición de niveles en un nivel de factor específico; y otro con la funcion step_downsample(), que es la encargada de crear una especificación de un paso de receta que eliminará filas de un conjunto de datos para hacer que la ocurrencia de niveles en un nivel de factor específico sea igual.

Para todo esto, utilizamos Workflows que es el componente fundamental que permite conectar el preprocesamiento de datos y el modelado predictivo dentro del marco de tidymodels. Para nuestro caso, usamos 2 Workflows ya que generamos 2 recetas para comparar cual es la mejor opción.

El modelo randomForest implementa el algoritmo de Breiman (basado en el código Fortran original de Breiman y Cutler) para la clasificación y la regresión. También puede utilizarse en modo no supervisado para evaluar las proximidades entre los puntos de datos.

A continuación, mostramos los valores obtenidos de los 2 modelos.

Como resultado se observan los siguientes valores del primer modelo:

	metric	estimator	mean	n	std_err
1	accuracy	multiclass	0.567	10	0.0103
2	roc_auc	hand_till	0.714	10	0.00698

Del segundo modelo se observan que los valores:

	metric	estimator	mean	n	std_err
1	accuracy	multiclass	0.495	10	0.0107
2	roc_auc	hand_till	0.683	10	0.00534

La precisión del modelo de aprendizaje automático es la medida utilizada para determinar qué modelo es el mejor para identificar las relaciones y los patrones entre las variables de un conjunto de datos basado en los datos de entrada. Cuanto mejor pueda generalizar un modelo a los datos “no vistos”, mejores predicciones y conocimientos podrá producir, lo que a su vez aportará más valor. Los resultados del primer modelo nos muestran una precisión más alta con respecto al segundo modelo.

La medida de ROC_AUC que es una medida de rendimiento para los problemas de clasificación. La ROC es una curva de probabilidad y la AUC representa el grado o la medida de separabilidad. Indica en qué medida el modelo es capaz de distinguir entre clases. Cuanto más alto sea el AUC, mejor será el modelo para predecir las clases 0 como 0 y las clases 1 como 1. Por analogía, cuanto más alto sea el AUC, mejor será el modelo para poder distinguir.

4. Conclusiones

En este trabajo se presentaron los primeros resultados iniciales de un proyecto de desarrollo en curso. Se presentaron dos tipos de metodologías para abordar el proyecto: con léxicos y con un clasificador basado en aprendizaje.

En el corpus detectamos una gran ambivalencia de las oraciones clasificadas como neutrales, ya que no siempre cumplen el requisito: tratar exclusivamente del big data y presentar información subjetiva. Esto se subsanará desarrollando la tarea vinculada con el análisis de objetos u aspectos, para así determinar si el objeto de la oración es efectivamente big data, en cuyo caso se puede analizar la polaridad de su contenido subjetivo. Esto nos permitirá generalizar el corpus, a cualquier oración que incluya el término “big data”.

Además, en la parte de aprendizaje conviene incluir una fase de “feature engineering”. También como desafío futuro, nos interesaría poder utilizar un sistema hibrido donde combinando lo mejor de cada método poder obtener un mejor resultado final.

Avanzar con este proyecto nos permitirá aportar una herramienta para distinguir una expresión “optimista” o “pesimista”, “promocional” o “crítica” del big data, permitiendo reintroducir esta diferencia en futuros trabajos lingüísticos o discursivos sobre el fenómeno, una necesidad señalada por la literatura (Paganoni, 2019). Este tipo de análisis es central en la indagación de sentido sociales y actitudes hacia fenómenos como el “big data”.

5. Bibliografía

Becerra, G. (2018). Big data como objeto de estudio y método para la investigación empírica en sociología y psicología social. *47 Jornadas Argentinas de Informática & Simposio Argentino de Tecnología y Sociedad*, 141–150. http://47jaiio.sadio.org.ar/sites/default/files/STS-13.pdf

Becerra, G., & López-alurralde, J. P. (2020). Hacia una exploración de las representaciones sociales en torno al big data. *49 Jornadas Argentinas de Informática & Simposio Argentino de Tecnología y Sociedad*.

Gravano, A., & Dell’Amerlina Ríos, M. (2014). *Spanish DAL: A Spanish Dictionary of Affect in Language*. http://digital.bl.fcen.uba.ar/Download/technicalreport/technicalreport_00001.pdf

Hvitfeldt, E., & Silge, J. (2021). *Supervised Machine Learning for Text Analysis in R*. https://smltar.com/

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage. https://doi.org/10.4135/9781473909472

Liu, B. (2015). *Sentiment Analysis In Computational Linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Paganoni, M. C. (2019). *Framing big data : a linguistic and discursive approach*. palgrave macmillan.

Straka, M., & Strakova, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencie*, 88–99.

Team R Core. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/