COYUNTURA DE LA AGRICULTURA EN ITALIA: METODOLOGÍA PARA UN ANÁLISIS TEXTUAL EN MEDIOS DIGITALES

Arenas-Estevez, Luisa Fernanda[1]
Rangel-Quiñonez, Henry Sebastián[2]

Recibido: 19/04/2025 Revisado: 23/06/2025 Aceptado: 10/03/2026

RESUMEN

Este estudio tiene como objetivo presentar una metodología replicable basada en técnicas de procesamiento de lenguaje natural (PLN) para el análisis automatizado del discurso mediático y aplicar dicha metodología al caso del sector agrario en Italia durante el primer semestre de 2024. Se analizaron 164 noticias de siete medios italianos con el fin de identificar los temas centrales, emociones predominantes y patrones discursivos presentes. Se emplearon herramientas de PLN en R para limpiar, *tokenizar* y vectorizar el corpus textual y luego se aplicaron modelos estadísticos como TF-IDF (acrónimo inglés de Frecuencia de término-Frecuencia inversa de documentos) y *Latent Dirichlet Allocation* (LDA) para identificar temáticas, además de utilizar el *NRC Emotion Lexicon* para clasificar emociones. El análisis temático basado en LDA identificó dos ejes centrales en la cobertura mediática: la gestión institucional del sector agroalimentario italiano y las protestas de los agricultores en la Unión Europea. El primer grupo refleja un enfoque sobre el desarrollo del sector, destacando temas como la innovación tecnológica y las políticas públicas, mientras que el segundo se centra en las manifestaciones de los agricultores y las respuestas institucionales. El análisis de sentimientos (*sentiment analysis*) reveló una predominancia de emociones como confianza (30%), anticipación (18%) y miedo (13%). Los picos de carga emocional negativa ocurrieron en los meses de febrero y marzo, asociados a crisis y protestas, mientras que las noticias más positivas se concentraron a finales de abril, coincidiendo con temas de innovación y cooperación internacional. Los resultados obtenidos contribuyen significativamente a la comprensión de cómo los medios construyen la narrativa sobre el sector agrario. Además, se destacan las emociones que influyen en la percepción pública, lo que puede ser de utilidad para encontrar referenciales de políticas en los medios de comunicación en este sector. Este estudio pone en evidencia el potencial del PLN para analizar discursos mediáticos, ofreciendo una base sólida para futuras investigaciones sobre la representación de la agricultura en los medios y su impacto en la opinión pública.

Palabras clave: agricultura, análisis de sentimientos, análisis de noticias, discursos mediáticos, procesamiento natural del lenguaje, Italia

[1] Ph.D. Fellow in Economics, Management and Accounting (Università degli Studi di Napoli Parthenope-UniParthenope, Italia); Máster avanzado en Economía y Política Agraria del Departamento de Agraria, Universidad Federico II-Unina, Italia); Magíster en Políticas Públicas (Universidad Nacional de Colombia-UNAL, Colombia); Economista (Universidad Industrial de Santander-UIS, Colombia). *Dirección postal:* Universidad de Nápoles, Parthenope. Via Generale Parisi, 13, 80132 Napoli (NA), Italia. Oficina 326. ORCID: https://orcid.org/0000-0001-5022-1854; *e-mail:* luisafernanda.arenasestevez001@studenti.uniparthenope.it

[2] Ph.D. Fellow in Economics, Management and Accounting (Università degli Studi di Napoli Parthenope-UniParthenope, Italia); Máster avanzado en Economía y Política Agraria del Departamento de Agraria, Universidad Federico I-Unina, Italia); Magíster en Estadística (Universidad Nacional de Colombia-UNAL, Colombia); Especialista en Estadística, Filósofo y Economista (Universidad Industrial de Santander-UIS, Colombia). Docente Asociado de la Facultad de Economía Universidad Santo Tomás (USTA), Bucaramanga. *Dirección postal:* Universidad Santo Tomás, Cra. 18 #9-27, Bucaramanga, Santander, Colombia. *ORCID:* https://orcid.org/0000-0002-6745-6753; *e-mail:* henry.rangel@ustabuca.edu.co

ABSTRACT

This study aims to present a replicable methodology based on natural language processing (NLP) techniques for the automated analysis of media discourse, and to apply this methodology to the case of the agricultural sector in Italy during the first half of 2024. A total of 164 news articles from seven Italian media outlets were analyzed to identify the central topics, predominant emotions, and discursive patterns. NLP tools in R were used to clean, tokenize, and vectorize the textual corpus, followed by the application of statistical models such as Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) to identify topics, and the NRC Emotion Lexicon to classify emotions. Based on LDA, the thematic analysis identified two central axes in the media coverage: the institutional management of the Italian agri-food sector and the protests of farmers in the European Union. The first group focuses on sector development, highlighting themes such as technological innovation and public policies, while the second centers on farmers' protests and institutional responses. The sentiment analysis revealed a predominance of emotions such as trust (30%), anticipation (18%), and fear (13%). Peaks of negative emotional load occurred in February and March, associated with crises and protests, while the most positive news was concentrated at the end of April, coinciding with topics of innovation and international cooperation. The results significantly contribute to understanding how the media constructs the narrative about the agricultural sector. Additionally, the emotions that influence public perception are highlighted, which can be useful for finding policy references in the media related to the agricultural sector. This study highlights the potential of NLP for analyzing media discourse, providing a solid foundation for future research on the representation of agriculture in the media and its impact on public opinion.

Key words: Agriculture, sentiment analysis, news analysis, media discourse, natural language processing, representations and imaginaries, Italy

RÉSUMÉ

Cette étude vise à présenter une méthodologie reproductible fondée sur des techniques de traitement automatique du langage naturel TALN (Traitement Automatique du Langage Naturel) pour l'analyse automatisée du discours médiatique, et à l'appliquer au cas du secteur agricole en Italie durant le premier semestre 2024. Un total de 164 articles issus de sept médias italiens ont été examinés afin d'identifier les thèmes centraux, les émotions prédominantes et les schémas discursifs présents. Des outils de TALN sous R ont été mobilisés pour nettoyer, tokeniser et vectoriser le corpus textuel, avant l'application de modèles statistiques tels que TF-IDF (Term Frequency-Inverse Document Frequency) et Latent Dirichlet Allocation (LDA) pour l'identification des thèmes, ainsi que du NRC Emotion Lexicon pour la classification des émotions. L'analyse thématique, fondée sur la LDA, a mis en évidence deux axes principaux dans la couverture médiatique : la gestion institutionnelle du secteur agroalimentaire italien et les mobilisations des agriculteurs au sein de l'Union européenne. Le premier axe reflète une approche orientée vers le développement du secteur, mettant en avant des thèmes tels que l'innovation technologique et les politiques publiques, tandis que le second se concentre sur les manifestations agricoles et les réponses institutionnelles qui y sont associées. L'analyse des sentiments révèle une prédominance d'émotions telles que la confiance (30 %), l'anticipation (18 %) et la peur (13 %). Les pics d'émotions négatives ont été observés en février et mars, en lien avec des crises et des mobilisations, tandis que les contenus les plus positifs se concentrent à la fin du mois d'avril, en coïncidence avec des thématiques liées à l'innovation et à la coopération internationale. Les résultats obtenus contribuent de manière significative à la compréhension des mécanismes par lesquels les médias construisent le discours relatif au secteur agricole. Par ailleurs, l'étude met en lumière les émotions susceptibles d'influencer la perception du public, ce qui peut s'avérer utile pour identifier des repères politiques dans la couverture médiatique de ce secteur. Elle souligne également le potentiel du NLP pour l'analyse des discours médiatiques, offrant ainsi une base solide pour de futures recherches sur la représentation de l'agriculture dans les médias et son impact sur l'opinion publique.

Mots-clés : agriculture, analyse des sentiments, analyse des actualités, discours médiatiques, traitement automatique du langage naturel, Italie

RESUMO

Este estudo tem como objetivo apresentar uma metodologia replicável baseada em técnicas de processamento de linguagem natural (PLN) para a análise automatizada do discurso midiático e aplicar essa metodologia ao caso do setor agrário na Itália durante o primeiro semestre de 2024. Foram analisadas 164 notícias de sete meios de comunicação italianos para identificar os temas centrais, emoções predominantes e padrões discursivos presentes. Ferramentas de PLN no R foram usadas para limpar, tokenizar e vetorizar o corpus textual, e, em seguida, foram aplicados modelos estatísticos como TF-IDF (frequência de termo, frequência inversa de documento TF-IDF, na sigla em inglês) e Latent Dirichlet Allocation (LDA) para identificar tópicos, além do NRC Emotion Lexicon para classificar emoções. A análise temática, baseada no LDA, identificou dois eixos centrais na cobertura midiática: a gestão institucional do setor agroalimentar italiano e os protestos dos agricultores na União Europeia. O primeiro grupo reflete um enfoque no desenvolvimento do setor, destacando temas como inovação tecnológica e políticas públicas, enquanto o segundo se concentra nas manifestações dos agricultores e nas respostas institucionais. A análise sentimental revelou uma predominância de emoções como confiança (30%), antecipação (18%) e medo (13%). Os picos de carga emocional negativa ocorreram entre fevereiro e março, associados a crises e protestos, enquanto as notícias mais positivas se concentraram no final de abril, coincidindo com temas de inovação e cooperação internacional. Os resultados obtidos contribuem significativamente para a compreensão de como os meios de comunicação constroem a narrativa sobre o setor agrícola. Além disso, destacam-se as emoções que influenciam a percepção pública, o que pode ser útil para encontrar referências de políticas nos meios de comunicação no setor agrário. Este estudo evidencia o potencial do PLN para analisar discursos midiáticos, oferecendo uma base sólida para futuras pesquisas sobre a representação da agricultura na mídia e seu impacto na opinião pública.
Palavras-chave: agricultura, análise sentimental, análise de notícias, discursos midiáticos, processamento de linguagem natural, representações e imaginários, Itália

1. INTRODUCCIÓN

El análisis de los discursos mediáticos permite explorar cómo se construyen los significados y qué emociones o valoraciones subyacen en el relato de las noticias. Los medios de comunicación no solo informan sobre los hechos, sino que también contribuyen a moldear las percepciones colectivas, reforzando o desafiando determinadas narrativas (Happer & Philo, 2013). Conceptos como *progreso*, *desarrollo* o *crisis* no son simplemente descripciones objetivas de fenómenos, sino construcciones cargadas de significado. Los medios de comunicación desempeñan un papel multifacético en la inclusión de las políticas agrarias y el sector agrícola en el discurso público. Actúan como transmisores de información, moldeadores de la opinión pública y como un vínculo entre los ciudadanos y los responsables de la toma de decisiones políticas en el sector agrícola. La forma en que los medios cubren estas temáticas, ya sea positiva o negativamente -y aquellas que eligen destacar- pueden tener implicaciones significativas para la dirección de las políticas agrarias y la percepción pública de la agricultura (Kr Yadav *et al.*, 2024; Mohr & Höhler, 2023).

En este sentido, comprender cómo ciertos temas son representados en los medios digitales de noticias resulta crucial. No obstante, el análisis de estos discursos enfrenta un desafío metodológico: la subjetividad inherente a la interpretación de grandes volúmenes de texto. Tradicionalmente los estudios cualitativos han abordado esta problemática mediante el análisis crítico del discurso y la hermenéutica, pero con la digitalización masiva de la información se han desarrollado nuevas metodologías que permiten un enfoque más estructurado. Entre ellas, el Procesamiento de Lenguaje Natural (PLN) y −en particular, el Análisis de sentimientos−, han cobrado relevancia como herramientas capaces de examinar el tono emocional de los textos y su polaridad (positiva, negativa o neutra) de manera sistemática y replicable (Cambria *et al.*, 2013; Liu, 2017; Pang & Lee, 2008). Estas técnicas han demostrado ser útiles para investigar cómo ciertos temas

son representados en los medios, proporcionando un enfoque cuantificable a fenómenos tradicionalmente abordados desde lo cualitativo.

Este artículo tiene un doble propósito. En primer lugar, busca acercar al lector al uso del Análisis de sentimientos como una herramienta metodológica accesible para el estudio de discursos mediáticos. Para ello, se presenta una guía detallada basada en herramientas computacionales de código abierto y uso libre, permitiendo su fácil replicación en diversos contextos. Con el fin de facilitar este proceso, se pone a disposición tanto el código en R como la base de datos utilizada, garantizando así la reproducibilidad del análisis[3]. En segundo lugar, se aplica esta metodología al estudio de noticias publicadas en siete periódicos digitales en Italia entre enero y junio de 2024, con el objetivo de identificar tendencias discursivas y emocionales en torno a la coyuntura del sector agrario del país. Además de ofrecer una aplicación concreta al caso de la agricultura en Italia, este enfoque resulta particularmente útil para investigadores que buscan analizar discursos en idiomas con los que no están familiarizados. El uso de técnicas automatizadas de PLN facilita la identificación de patrones discursivos sin requerir un dominio experto del idioma analizado, ampliando así las posibilidades de estudio en contextos lingüísticos diversos.

La elección de Italia como país de análisis se fundamenta en su posición estratégica dentro +del sector agrario europeo, al ser uno de los principales productores y exportadores del continente, así como el país con mayor cantidad de productos con Denominación de Origen Protegida (DOP) e Indicación Geográfica Protegida (IGP) (Istat, 2024). Durante el periodo analizado, de enero a junio de 2024, se registraron en Italia y otros países europeos protestas de agricultores que cuestionaban políticas ambientales de la Unión Europea, el incremento de los costos de producción y acuerdos comerciales percibidos como desfavorables. Estas manifestaciones, ampliamente cubiertas por los medios italianos, crearon un contexto propicio para explorar cómo los discursos mediáticos construyen la percepción pública del conflicto agrario.

2. METODOLOGÍA Y MATERIALES

Para el estudio se recopilaron noticias de siete periódicos italianos de acceso libre en la web, a saber: Avvenire, Corriere della Sera, Il Fatto Quotidiano, Il Sole 24 Ore, Il Messaggero, La Repubblica y La Stampa. El periodo de análisis abarcó desde enero hasta junio de 2024, en tanto que la recolección de datos se realizó en dos fechas específicas: el 16 de mayo y el 26 de junio. La búsqueda de artículos se llevó a cabo a través de los motores de búsqueda internos de cada periódico, empleando como palabra clave *agricoltura*. De cada noticia seleccionada se extrajo la información del nombre del periódico (Fuente), título de la noticia (Título), fecha de publicación (Fecha), los primeros párrafos de la noticia (Texto) y enlace al artículo (Vínculo). Estos datos constituyen la base del análisis textual. El procesamiento y análisis de los datos se realizó en el lenguaje de programación *R* (R Core Team, 2022), utilizando paquetes especializados para el análisis de texto.

2.1. LIMPIEZA DE DATOS

El vector Texto constituye el núcleo del análisis textual en este estudio. Antes de proceder con el procesamiento del contenido fue necesario realizar una limpieza preliminar del texto con el fin de garantizar resultados más precisos y consistentes. Para ello se emplearon diversos paquetes del entorno R, entre los que destacan *dplyr* (Wickham *et al*., 2023) *tidytext* (Silge & Robinson, 2016), y *tm* (Feinerer *et al*., 2008; Feinerer & Hornik, 2024). El proceso de limpieza consistió en la eliminación de signos ortográficos, como apóstrofes, acentos, comas y puntos, así como de símbolos de escasa relevancia analítica, incluyendo números y palabras vacías (*stopwords*), tales como artículos, preposiciones y pronombres (Sarica & Luo, 2021).

Para llevar a cabo la limpieza textual se utilizaron expresiones regulares (*Regex*), un sublenguaje especializado en la búsqueda y manipulación de patrones dentro de cadenas

[3] El código y los datos utilizados para el procesamiento y análisis del texto están disponibles en el repositorio: https://github.com/sebasrangel29/nlp_agricoltura

de texto (Aho, 1990; Fitzgerald, 2012). *Regex* permite establecer reglas específicas para identificar, reemplazar o eliminar caracteres no deseados, facilitando así la normalización del contenido textual. En este estudio, se aplicaron expresiones regulares en R para eliminar signos especiales como apóstrofes, comillas y signos de interrogación, así como los acentos graves y agudos típicos del italiano (è, à, ì, ò, ù, é). También se eliminaron números y otros símbolos que podrían interferir en el análisis posterior, con el objetivo de estandarizar el texto y garantizar su consistencia.

El corpus de *stopwords* en italiano proviene del paquete *tm* y contiene 274 palabras, las cuales pueden ser personalizadas con el comando *add_row()*. Sin embargo, el nivel de limpieza depende del idioma y del contexto de estudio ya que, en algunos casos, ciertos elementos pueden ser relevantes. Por ejemplo, los números pueden ser fundamentales en estudios que analizan valoraciones cuantitativas en respuestas abiertas, mientras que los pronombres pueden ser clave en investigaciones sobre discurso y género.

A continuación, se muestra un fragmento del código utilizado en R para limpiar el texto del *data frame* (DF):

```
#Eliminación de caracteres no deseados (Regex)
DF$Texto <- gsub(pattern =»\\’», replacement = « «, Texto) # Reemplaza apóstrofes por espacio
DF$Texto <- gsub(«[^[:alpha:] ]», «», Texto) # Elimina cualquier símbolo que no sea letra
DF$Texto <- gsub(«[ÈàèìòùéÌ]», «», Texto) # Elimina acentos italianos

# Cargar lista de stopwords en italiano
italian <- data.frame(word = stopwords («italian»))

# Agregar palabras personalizadas
custom_stopwords <- italian %>%
 add_row(word = «ue») %>%
 add_row(word = «de») %>%
 add_row(word = «cos») # Continuar con otras palabras si es necesario
# Eliminación de stopwords del texto
DF_SS <- DF %>%
 unnest_tokens(word, Texto) %>%
 anti_join(custom_stopwords, by = «word»))
```

Tras la aplicación de estos pasos, el texto original:

> Da Bruxelles sono arrivate risposte alle istanze degli agricoltori. Le manifestazioni di piazza sono state importanti. Però attenzione: le proteste hanno accelerato le decisioni ma le soluzioni non sono venute dalla piazza ma dal lavoro e dal confronto impostato nei mesi scorsi con le istituzioni e con gli altri governi. Ne è convinto il presidente della Coldiretti Ettore Prandini che a oltre due mesi dalle prime manifestazioni di protesta traccia un primo bilancio delle risposte messe in campo da Bruxelles

Se transforma en un texto limpio:

> bruxelles arrivate risposte istanze manifestazioni piazza state importanti attenzione proteste accelerato decisioni soluzioni venute piazza lavoro confronto impostato mesi scorsi istituzioni altri governi convinto presidente coldiretti ettore prandini oltre mesi prime manifestazioni protesta traccia primo bilancio risposte messe campo bruxelles

Para concluir el preprocesamiento del texto se realizó la extracción del *Stem* (raíz) de cada palabra con el paquete SnowballC (Bouchet-Valat, 2023). El *Stemming* reduce las palabras a su forma base, eliminando variaciones morfológicas que podrían afectar el análisis. Por ejemplo, en italiano, términos como *manifestazioni* y *manifestazione* pueden reducirse a *manifestazion*, lo que ayuda a agrupar términos semánticamente equivalentes en un mismo análisis. Sin embargo, su uso no es siempre es recomendable, ya que la reducción excesiva de palabras puede generar pérdidas de significado o ambigüedades en ciertos contextos.

El código en R para realizar el *stemming* es el siguiente:

```
# Aplicar stemming en italiano
DF_stemmed <- DF_SS %>%
 mutate(word = wordStem(word, language = «italian»))
```

Este último proceso también implica la *tokenización*, es decir, la segmentación del texto en unidades básicas de análisis. La *tokenización* es crucial en el procesamiento del lenguaje natural, ya que convierte el texto en elementos estructurados que pueden ser analizados computacionalmente (Grefenstette, 1999).

A continuación, se muestra cómo algunas palabras se transforman tras la aplicación del *stemming*:

Tabla 1
Ejemplo tokens con stem

Original	stem
Original	stem
arrivate	arriv
risposte	rispost
agricoltori	agricoltor
decisioni	decision

Aunque este estudio emplea una *tokenización* basada en unigramas (palabras individuales), es relevante señalar que existen alternativas metodológicas que pueden enriquecer el análisis textual, como la *tokenización* por bigramas o n-gramas, que permite capturar expresiones compuestas (Silge & Robinson, 2017). Asimismo, otra estrategia avanzada es la aplicación del etiquetado gramatical o POS tagging (*Part-of-Speech Tagging*), que consiste en asignar a cada palabra una categoría sintáctica (sustantivo, verbo, adjetivo, etc.). Esta clasificación permite filtrar palabras según su función gramatical, priorizar categorías relevantes para el análisis -por ejemplo, sustantivos para extraer temas o adjetivos para identificar valoraciones- y estudiar estructuras lingüísticas complejas (Straka & Straková, 2017).

2.2. MODELIZACIÓN TEXTUAL

Tras la limpieza y *tokenización* del texto, se construye una matriz dispersa (*Sparse Matrix*) (Qaiser & Ali, 2018), que permite representar numéricamente la frecuencia de las palabras en cada noticia. Esta matriz contiene tantas filas (n) como noticias analizadas (164) y, por columnas (m), el total de distintas palabras encontradas: 3.473 si se trabaja con el *Data Frame* (DF) sin *stopwords*, o 2.514 si se trabaja solamente con las *raíces*. Cada elemento x_{ij} de la matriz indica la cantidad de veces que la palabra j aparece en la noticia i.

A continuación se presenta un ejemplo simplificado de una matriz dispersa. En este caso, la Figura 1 ilustra una versión reducida con 6 filas –simulando noticias– y 6 columnas –representando diferentes palabras identificadas en el corpus–.

Figura 1
Matriz dispersa

1	0	0	1	0	0
0	0	3	0	2	1
1	2	0	1	0	0
1	0	1	0	0	0
0	1	0	2	0	0
1	0	0	0	0	1

A partir de la matriz dispersa se calcularon dos medidas importantes para el análisis textual: en primer lugar, la frecuencia de término (*Term Frequency, TF*), que mide cuántas veces aparece una palabra en un documento en relación con la longitud del documento y que -en este caso- representa la cantidad de veces que un término aparece en una noticia en relación con su cantidad total de palabras. Se calcula como:

$$TF_{ij} = \frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}}$$

donde x_{ij} es el número de veces que la palabra j aparece en la noticia i y el denominador es la suma total de palabras en esa noticia.

Por su parte, la segunda medida es la frecuencia inversa de documento (*Inverse Document Frequency*, IDF), que evalúa la importancia de una palabra considerando en cuántas noticias aparece. Si un término se encuentra en muchas noticias, su peso se reduce, ya que se considera menos relevante para diferenciar documentos. Su cálculo es el siguiente:

$$IDF_j = \log\left(\frac{N}{n_t}\right)$$

donde *N* es el número total de noticias y n_t es el número de noticias en los que aparece la palabra j.

El producto de ambas medidas da como resultado el TF-IDF (*Term Frequency-Inverse Document Frequency*), que cuantifica la importancia de una palabra en una noticia dentro del conjunto total de textos. Este valor aumenta proporcionalmente con el número de veces que un término aparece en una noticia, pero se equilibra con su presencia en el resto del corpus. Su fórmula es:

$$TF\_IDF_{ij} = TF_{ij} * IDF_j$$

Este procedimiento permite convertir el texto en datos cuantificables y facilita su posterior análisis mediante técnicas estadísticas. Un método ampliamente utilizado en este contexto es *Latent Dirichlet Allocation* (LDA) diseñado para extraer patrones en grandes volúmenes de texto y clasificar términos en distintos temas de manera probabilística (Blei *et al.*, 2003). LDA se basa en la idea de que los textos pueden descomponerse en distribuciones latentes de temas (k), permitiendo identificar patrones semánticos sin necesidad de etiquetado previo. En este estudio, se seleccionó k = 2 tópicos siguiendo un criterio interpretativo, con el objetivo de privilegiar la claridad temática y la utilidad descriptiva de los resultados. No obstante, existen métricas cuantitativas que pueden ser útiles para determinar el número óptimo de tópicos (Gan & Qi, 2021).

En este modelo la asignación de palabras a temas se realiza de manera probabilística, lo que facilita el descubrimiento de estructuras subyacentes en grandes volúmenes de texto. Su aplicación es útil para el análisis exploratorio de noticias, porque permite identificar de manera automática los principales tópicos abordados en el corpus, proporcionando una representación estructurada del contenido textual.

A continuación se presenta el código en R utilizado para calcular TF, IDF, TF-IDF y aplicar el modelo LDA:

```
# Construcción de la matriz de términos
matrix1 <- DF_stemmed %>%
 count(ID, word) %>%
 cast_dtm(document = ID, term = word,
value = n, weighting = tm::weightTf)

# Aplicación del modelo LDA con método
Gibbs
lda <- LDA(matrix1, k = 2, method = 'Gibbs',
control = list(seed = 1111))
```

En este ejercicio para la representación numérica del corpus textual se optó por la codificación TF-IDF, dado su balance entre simplicidad, interpretabilidad y capacidad para identificar términos relevantes dentro de un conjunto de documentos. Esta elección metodológica se ajusta al objetivo del estudio, centrado en identificar patrones temáticos diferenciadores dentro del discurso mediático. Si bien existen alternativas más recientes como los modelos de *word embeddings* (por ejemplo, Word2Vec o GloVe), capaces de capturar relaciones semánticas más complejas según el contexto del n-grama, su aplicación requiere entrenamiento en corpus mucho más extensos. En este caso, el uso de TF-IDF resulta más adecuado por su compatibilidad con el modelo LDA y su facilidad de interpretación en contextos exploratorios.

2.3. ANÁLISIS DE SENTIMIENTOS

El análisis de sentimientos es una técnica ampliamente utilizada en el campo del PLN, su propósito es identificar y clasificar la carga emocional contenida en un texto (Cambria *et al.*, 2013; Liu, 2017). Esta metodología permite determinar si un mensaje transmite una valoración positiva, negativa o neutra, así como detectar emociones subyacentes que ayudan a

caracterizar el tono del discurso. En el contexto del análisis de noticias sobre agricultura en Italia, el uso del análisis de sentimientos proporciona una aproximación al juicio de valor implícito en los textos periodísticos, permitiendo identificar tendencias discursivas y posibles sesgos en la representación del sector agrario.

El análisis sentimental se realizó sobre un vector de texto previamente procesado y depurado de *stopwords*, compuesto por un total de 6793 palabras. A cada término se le asignó un puntaje emocional según su carga semántica, dentro de una escala discreta que permite clasificar las palabras en distintas categorías afectivas. Esta asignación se basa en lexicones de sentimientos elaborados por lingüistas y especialistas en PLN. Entre los más utilizados se encuentran los desarrollados por el paquete *syuzhet* (Jockers, 2015), el lexicón de Bing Liu (Liu, 2012) y el *NRC Emotion Lexicon* (Mohammad, 2020; Mohammad & Turney, 2013) todos implementados en *R*.

En este estudio se optó por el lexicón NRC, que incluye 6.468 términos en varios idiomas, entre ellos el italiano. Este recurso permite realizar una doble clasificación: por un lado, identifica palabras como positivas o negativas; y, por otro, las categoriza según las ocho emociones básicas de Plutchik (1980): ira, anticipación, disgusto, miedo, alegría, tristeza, sorpresa y confianza. Es importante señalar que algunas palabras pueden pertenecer simultáneamente a más de una categoría emocional, por lo que el corpus final asciende a 13.901 entradas.

A continuación se presenta el código en *R* utilizado para generar el corpus en italiano y aplicarlo al vector de noticias:

```
# Cargar lexicón NRC en italiano
sentiments
<- get_nrc_sentiment(DF_SS$word,
lang="italian")
```

Este proceso permite cuantificar las emociones expresadas en las noticias, facilitando el análisis de los patrones sentimentales y las tendencias discursivas. Cada una de las emociones identificadas por el lexicón NRC tiene un significado particular que orienta su interpretación en el análisis del discurso mediático. Por ejemplo, siguiendo a Plutchik (1980) la **alegría** está vinculada con sensaciones de felicidad, placer y satisfacción, lo que indica la presencia de interacciones o situaciones gratificantes. **Tristeza** –en cambio– suele representar pérdidas, decepción o dolor y puede reflejar una narrativa marcada por el desánimo o la frustración. La **ira** aparece como respuesta emocional frente a situaciones de injusticia o conflicto, manifestándose a través de un lenguaje más agresivo o confrontativo.

Por su parte, el **Miedo** señala percepciones de inseguridad o amenaza, anticipando eventos negativos o riesgos potenciales. La **confianza** expresa seguridad y aceptación, reflejando relaciones positivas o estabilidad institucional. El **disgusto** se vincula con sentimientos de rechazo hacia aspectos percibidos como indeseables o poco apropiados, ya sea por razones morales, estéticas o de calidad. **Sorpresa** responde a eventos inesperados, que pueden ser tanto positivos como negativos, mientras que **anticipación** remite a expectativas y esperanzas hacia el futuro, capturando el deseo o la curiosidad por lo que está por venir. Estas categorías permiten no solo cuantificar emociones, sino también interpretar el tono subyacente con el que se narran los acontecimientos relacionados con el sector agrícola italiano.

El enfoque de análisis de sentimientos adoptado en este estudio se basa en el uso de un lexicón de emociones, conocido como *lexicon-based,* el cual ofrece ventajas como la transparencia y la facilidad de implementación. No obstante, existen enfoques más avanzados basados en modelos de lenguaje de gran escala (LLMs), como BERT (*Bidirectional Encoder Representations from Transformers*), que permiten captar matices semánticos más complejos y realizar una clasificación contextualizada del sentimiento. Estos métodos, si bien más potentes, requieren grandes volúmenes de datos etiquetados para su entrenamiento, o el uso de modelos previamente entrenados y adaptados al idioma específico.

3. RESULTADOS

3.1. DESCRIPCIÓN DE DATOS

El corpus analizado estuvo compuesto por 164 noticias publicadas entre el 8 de enero y el 24

de junio de 2024. La distribución temporal de las publicaciones revela una mayor concentración durante los meses de febrero y marzo, los cuales agruparon el 55,4 % del total de noticias recopiladas. Esta concentración podría estar relacionada con eventos específicos o con una mayor atención mediática al sector agrícola en ese periodo. En cuanto a la distribución por fuente, la mayoría de los artículos provienen de los periódicos *Il Sole 24 Ore* y *Il Fatto Quotidiano*, que -en conjunto- representan el 65,3% del total. En los casos de *La Repubblica*, *Il Messaggero* y *Corriere Della Sera* se tuvo la limitación de no poseer una suscripción, lo cual redujo el número total de consultas gratuitas (Tabla 2).

Tabla 2
Número de noticias por fuente

Fuente	n	%
Avvenire	11	6,71
Corriere Della Sera	4	2,44
Il Fatto Quotidiano	40	24,39
Il Messaggero	3	1,83
Il Sole	67	40,85
La Repubblica	2	1,22
La Stampa	37	22,56
Total	**164**	**100**

El siguiente paso consistió en explorar el contenido de las noticias mediante el análisis de frecuencias de palabras. Esta etapa permite identificar los términos más recurrentes y, por ende, los ejes temáticos que predominan en el discurso periodístico sobre el sector agrícola italiano. La Figura 2 presenta una visualización del corpus final en forma de nube de palabras, donde el tamaño de cada término es proporcional a su frecuencia absoluta. Con el objetivo de evitar sesgos derivados del proceso de recolección, basado en la búsqueda por la palabra clave *agricultura*, se eliminaron intencionalmente del corpus los términos directamente relacionados con dicho concepto, como *agricoltura*, *agricola* y variantes asociadas. Esta decisión metodológica buscaba resaltar otros elementos del discurso periodístico que podrían pasar desapercibidos frente al peso predominante del término clave. Así, entre las palabras con mayor frecuencia destacan *trattori*, *ministro*, *settore*, *consiglio*, *europea* y *protesta*, lo que permite anticipar algunos de los temas más relevantes en el tratamiento mediático de la coyuntura agraria.

Entre los términos destacados en frecuencia llaman especialmente la atención las palabras *trattori* y *protesta*, que remiten a una de las coyunturas más relevantes del sector agrícola

Figura 2
Nube de frecuencia

italiano durante el primer semestre: las manifestaciones de agricultores y el uso simbólico del tractor como elemento de protesta. Por ello, se realiza un análisis más detallado sobre la distribución de estas palabras según la fuente periodística, en el corpus sin *Stem*. En el caso de *protest*, el periódico Il *Fatto Quotidiano* registró 26 menciones, lo que equivale a un promedio de 0,65 menciones por noticia. Le sigue *Il Sole 24* Ore con 9 apariciones, 0,13 menciones por noticia. Por su parte, la palabra *trattor* aparece 17 veces en *Il Fatto Quotidiano*, 0,43 menciones por noticia y 7 veces en *La Stampa*, 0,19 menciones por noticia. Estos datos sugieren que *Il Fatto Quotidiano* fue el medio con mayor cobertura discursiva de las llamadas «protestas de los tractores».

Además del análisis de frecuencia se calculó una matriz de similitud utilizando el índice TF-IDF, implementado a través del paquete *widyr* (Silge & Robinson, 2016), con el objetivo de comparar la cercanía temática entre los textos. Esta matriz permite cuantificar el grado de similitud entre pares de noticias, identificando aquellas que comparten un contenido muy similar. Esta técnica resulta especialmente útil para detectar noticias duplicadas o *recicladas*, es decir, textos que han sido republicados con mínimas variaciones. En este estudio no se encontraron coincidencias con un nivel de similitud superior al 34/ %, lo que sugiere que el corpus está compuesto por textos únicos y sin repeticiones significativas.

Tabla 3
Grado de similitud entre noticias

Ítem 1	Ítem 2	Similitud
21	90	0,34
90	21	0,34
31	102	0,26
102	31	0,26
72	73	0,25
73	72	0,25
151	154	0,24
154	151	0,24

3.2. MODELO DE CLASIFICACIÓN

Tras aplicar el modelo LDA al corpus sin *stem*, se identificaron dos grupos temáticos principales, cada uno caracterizado por un conjunto específico de palabras con alta frecuencia relativa. El grupo 1 está conformado por términos como *europe, ministr, protest, manifest, trattor, govern, consigl, politic,* y *commission*, los cuales apuntan a un eje temático centrado en las manifestaciones agrícolas y su repercusión política e institucional. Las palabras *giorn (*días*)* y *arriv (*llegada*)* sugieren una narrativa temporal o dinámica de los eventos, mientras que *part (*partidos*)* y *paes (*país*)* indican una dimensión nacional y política del conflicto. Este grupo refleja el momento de tensión entre los agricultores y las autoridades europeas e italianas, destacando las movilizaciones de tractores, las protestas y la cobertura en torno a la respuesta gubernamental y de la Comisión Europea. La Figura 3 muestra en orden descendente el aporte de las principales palabras del grupo 1.

En el Grupo 2, por su parte, predominan términos como *settor, stat, italian, produzion, aliment, tecnolog, milion, innov* y *svilupp*, que delinean un eje más productivo, económico y técnico. Este grupo se enfoca en el desarrollo del sector agroalimentario en Italia, incluyendo elementos como la innovación tecnológica, el financiamiento (*milion*), el rol del Estado (*stat*) y el discurso sobre el crecimiento del sector (*svilupp, nuov, quasi*). Es una narrativa centrada en el fortalecimiento del sector, resaltando el papel del Estado, las políticas públicas y las dinámicas de innovación territorial.

De otro lado, el análisis de tópicos permitió identificar qué dimensiones del sector del agro están siendo priorizadas en la cobertura mediática, lo cual reviste especial importancia dado que los medios no solo informan, sino que también construyen marcos interpretativos que influyen en la opinión pública, las prioridades gubernamentales y la legitimidad de determinadas políticas (Baker & Irani, 2014; Haller *et al*., 2019; Kr Yadav., 2024; Mohr & Höhler, 2023). En el caso analizado, la emergencia de un bloque discursivo centrado en las protestas sociales y la respuesta institucional revela cómo la agricultura es presentada como un campo de conflicto. Comprender cómo se articulan estas tensiones permite anticipar resistencias sociales y gestionar la comunicación institucional. El modelado temático facilita la detección de

Figura 3
Palabras del grupo 1

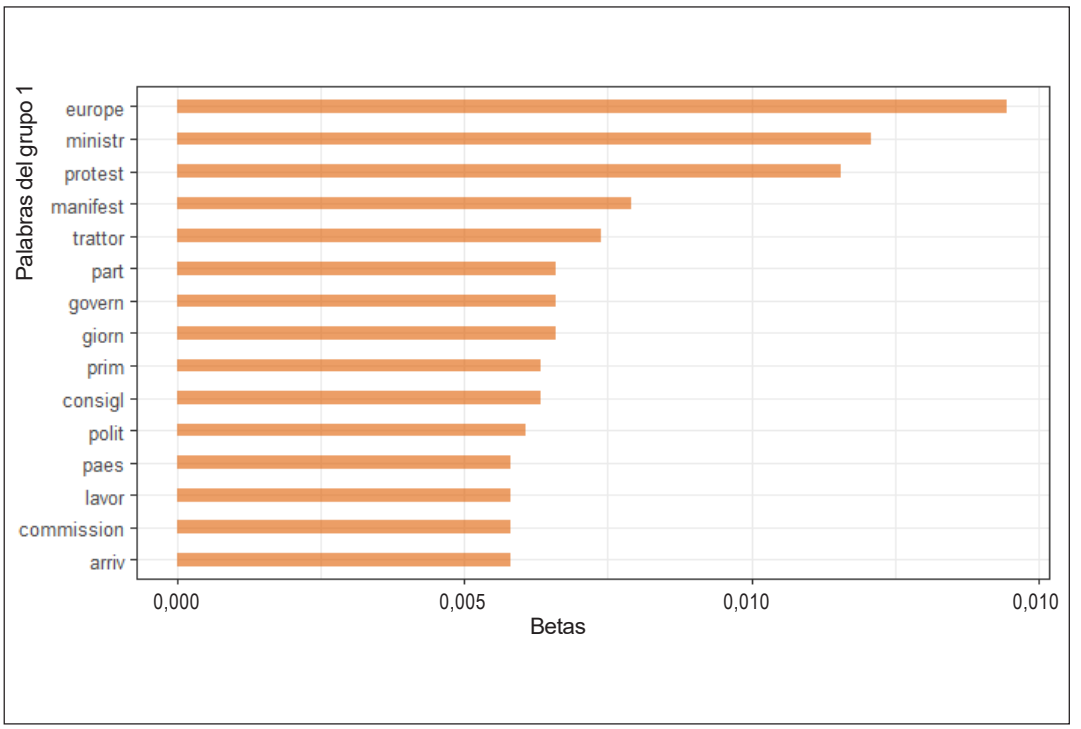
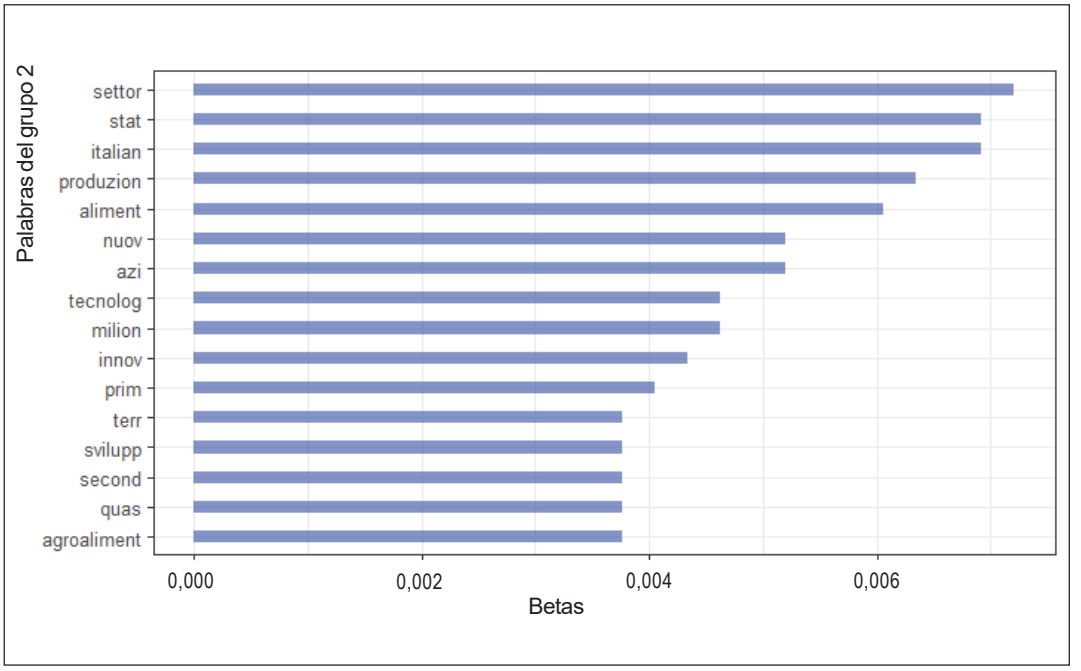


Figura 4
Palabras del grupo 2



narrativas emergentes, como la aparición progresiva de términos vinculados a sostenibilidad, innovación o digitalización, que revelan el desplazamiento en las prioridades discursivas hacia enfoques más tecnológicos y ambientales. En conjunto, este análisis de tópicos ofrece una base empírica para identificar la demanda y la percepción social del sector agro.

3.3. ANÁLISIS DE SENTIMIENTOS

Tal como se ha señalado previamente, el análisis de sentimientos permite evaluar la carga emocional contenida en los textos periodísticos, ofreciendo una visión más profunda sobre el tono y la perspectiva con la que los medios abordan el tema agrícola. Esta técnica resulta especialmente útil para detectar patrones de representación emocional, así como para analizar el impacto simbólico del discurso mediático. La Figura 5 presenta la distribución general de emociones asociadas a las palabras contenidas en las noticias. Se observa que el 30% de los términos están vinculados con la emoción de confianza, seguida por anticipación (18%) y miedo (13%). Esta prevalencia de emociones como la confianza y la anticipación sugiere un enfoque relativamente constructivo en la cobertura mediática, aunque también se evidencia una presencia de emociones negativas, asociado posiblemente a los momentos de mayor conflicto.

El análisis también permite identificar ejemplos concretos de palabras asociadas a cada emoción. La Figura 6, muestra cuatro de las emociones predominantes en las noticias y algunas de las palabras relacionadas. Por ejemplo, *tutela*, *accordo*, *presidente*, *legge*, *fatto* y *burocrazia* fueron clasificadas dentro de la categoría **confianza**. En contraste, términos como *problema*, *brutale*, *strage*, *allarme* y *manifestazione* se vinculan con **miedo.**

Para observar cómo varía el tono emocional de las noticias a lo largo del tiempo, se organizó el corpus cronológicamente según la fecha de publicación. A diferencia de los análisis anteriores basados en palabras individuales, en este caso cada noticia fue considerada como una unidad completa. A cada una se le asignó un puntaje total sumando +1 por cada palabra positiva y -1 por cada palabra negativa. De esta forma cada noticia

Figura 5
Clasificación de las emociones según el NRC

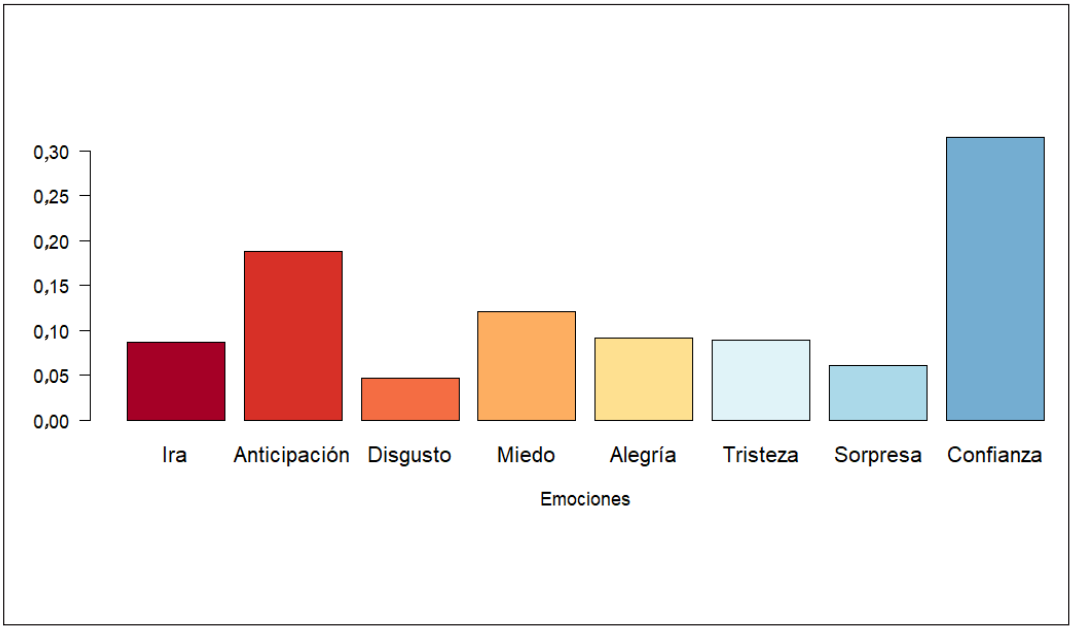
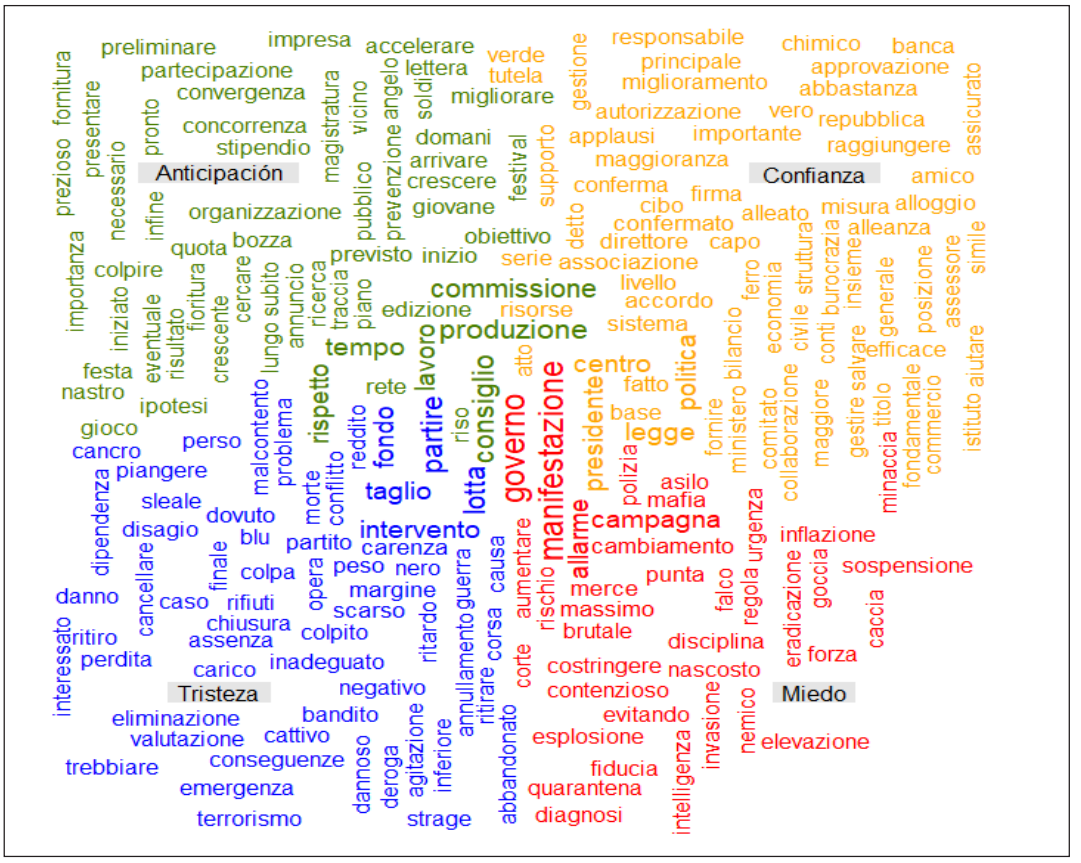


Figura 6
Muestra de palabras según la emoción



tenía un puntaje igual a la sumatoria de los valores obtenidos por las palabras que lo componen. Este valor fue posteriormente escalado entre -1 (carga completamente negativa) a +1 (carga completamente positiva) a través de la siguiente ecuación:

$$reescala_i = 2 * \frac{Puntaje\ total_i - Min(Puntaje\ total_i)}{Max(Puntaje\ total_i) - Min(Puntaje\ total_i)} - 1.$$

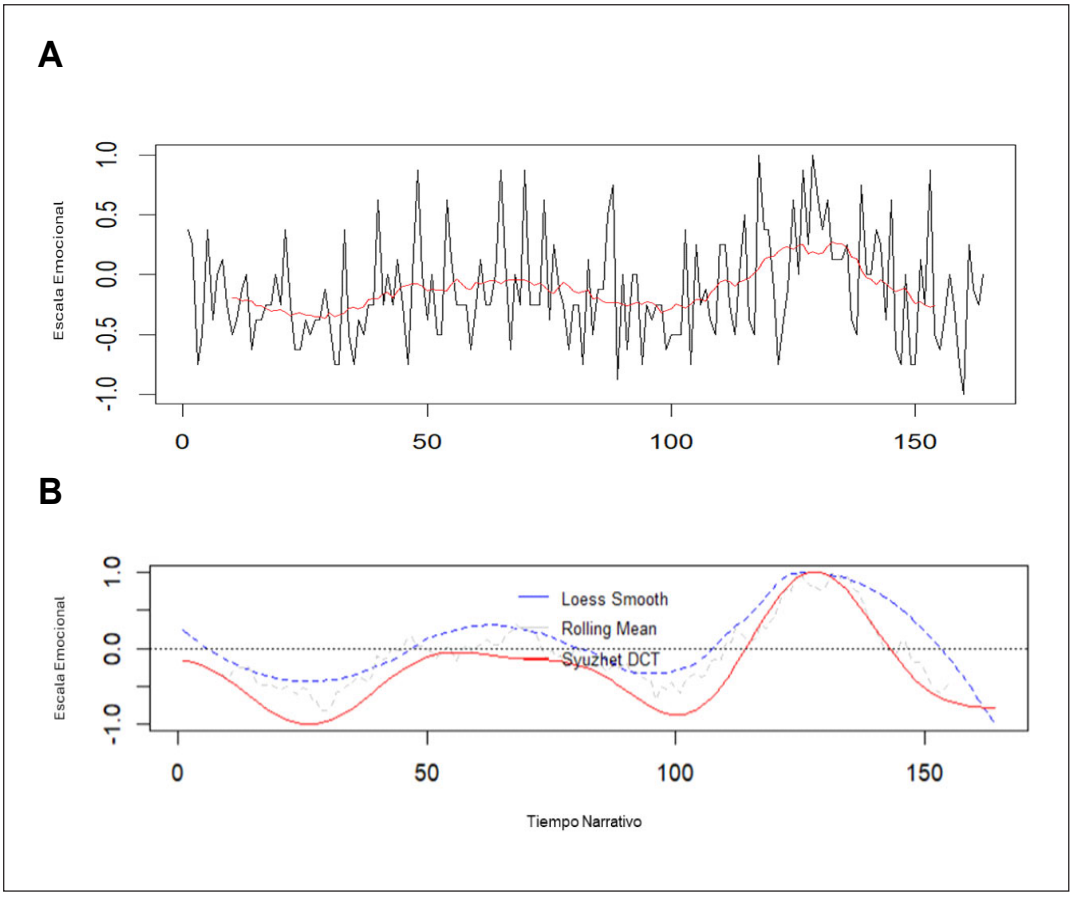
La dinámica emocional resultante se representa en las Figuras 7a y 7b. En ambas, el eje horizontal corresponde al *tiempo narrativo*, es decir, las noticias ordenadas cronológicamente desde la número 1 hasta la 164, mientras que el eje vertical muestra el puntaje de sentimiento escalado.

La Figura 7a fue generada manualmente aplicando la fórmula de escalamiento, lo que permitió graficar la secuencia de emociones (línea negra) y añadirle un suavizamiento con un promedio móvil (línea roja) calculado mediante una ventana de 20 noticias, lo que facilita la observación de las tendencias generales como los posibles ciclos. Por su parte, la Figura 7b fue elaborada con la función *simple_plot* del paquete *syuzhet*, que incluye también una línea de promedio móvil, junto con dos curvas suavizadas: una mediante el método LOESS (*Locally Estimated Scatterplot Smoothing*) y otra a través de una Transformada Discreta del Coseno (DCT, por sus siglas en inglés) (Jockers, 2023).Las representaciones suavizadas permiten observar cómo varía la carga emocional promedio de las noticias durante el periodo analizado, de enero a junio de 2024. En ambos gráficos se aprecia una carga

predominantemente negativa al inicio del año, con caídas significativas en las noticias número 32 y 99. La noticia 32, publicada a mediados de febrero, llevaba por título: *«El rugido de 200 tractores en Asti: 'La agricultura se muere, no hay futuro para nuestros hijos'»*. La segunda, publicada a finales de marzo, tenía como titular: *«Alarma en la agricultura: la extracción de agua de los ríos se reducirá a la mitad en 2025»*. A partir de ese momento se observa una recuperación progresiva del tono emocional, que alcanzaría su punto más alto hacia finales de abril, alrededor de la noticia número 129, titulada: *«Uganda, agricultura digital y formación para invertir en los jóvenes»* (traducción propia). Cabe señalar que los titulares presentados corresponden a traducciones propias del original en italiano.

El análisis de sentimientos proporciona una capa interpretativa adicional que permite evaluar no solo qué se dice del sector agrario, sino cómo se dice. Identificar los tonos emocionales predominantes, como la anticipación, el miedo o la confianza, resulta útil para entender el clima social que enmarca la recepción de determinados temas o políticas. Por ejemplo, picos de carga emocional negativa coincidentes con momentos de protesta o conflicto pueden indicar umbrales críticos de tensión social, cuya gestión requiere estrategias de comunicación más empáticas y políticas de contención específicas. Asimismo, la evolución temporal del sentimiento permite monitorear el impacto emocional de decisiones gubernamentales, cambios normativos o eventos climáticos sobre

Figura 7
Escala emocional vs. tiempo narrativo



la narrativa pública. Esta información resulta clave para los actores institucionales y productivos, ya que facilita un diagnóstico más fino del estado del sector, orienta la gestión del riesgo reputacional y permite diseñar intervenciones más alineadas con las percepciones sociales y los niveles de aceptación pública.

4. CONCLUSIONES

El presente artículo propone una metodología para el análisis del discurso periodístico en el ámbito agrícola, reconociendo la influencia que los medios de comunicación ejercen sobre la agenda pública y la opinión ciudadana. El aporte central de este estudio consiste en documentar una estrategia reproducible y objetiva, para identificar tanto los ejes temáticos como la carga emocional de las noticias, a partir de un corpus compuesto de 164 noticias que contienen la palabra *agricoltura*, publicadas durante el primer semestre de 2024 en siete diarios italianos durante un periodo marcado por intensas protestas de agricultores. Este documento se acompaña del material necesario para su replicación: se incluye la matriz de noticias recolectadas y el código comentado, complementando así la descripción metodológica presentada. Para facilitar su aplicación se empleó el lenguaje de programación R, una herramienta de uso libre. En conjunto, esta metodología demuestra cómo el análisis automatizado del lenguaje puede contribuir significativamente al estudio de fenómenos sociales, permitiendo una estandarización metodológica y una exploración profunda de los discursos mediáticos.

Los resultados del modelo de tópicos LDA permiten observar dos núcleos temáticos en las noticias: i) el primero, vinculado a las movilizaciones agrícolas y las respuestas institucionales, y, ii) el segundo, orientado a los aspectos estructurales y técnicos del sistema agroalimentario italiano. Esta distinción muestra cómo la prensa representó la agricultura como un espacio de conflicto sociopolítico y como un motor económico e innovador del país. El análisis de sentimientos aportó una dimensión adicional sobre el tono emocional de las noticias. Las emociones predominantes fueron confianza, anticipación y miedo, evidenciando una narrativa atravesada por la esperanza institucional, la expectativa frente a políticas futuras y la preocupación ante los desafíos del sector. El seguimiento cronológico del sentimiento identificó puntos críticos vinculados a eventos de protesta y crisis hídrica, así como momentos de recuperación emocional ligados a noticias de innovación y cooperación internacional.

En conjunto, este estudio pone de relieve el potencial del procesamiento de lenguaje natural y del análisis de sentimientos como herramientas para el estudio empírico del discurso mediático. Más allá de la cuantificación de términos, estas técnicas permiten acceder a las dimensiones simbólicas y emocionales con las que se construyen narrativas sobre el mundo agrícola. Los resultados pueden servir de base para investigaciones futuras sobre la representación mediática del sector agrario, el papel de los medios en la formación de la opinión pública y las tensiones entre el campo y las políticas.

Una de las limitaciones metodológicas del estudio radica en la distribución desigual del corpus entre las fuentes analizadas, como se detalla en la Tabla 2. La concentración de noticias en algunos medios –especialmente– se explica en parte por la falta de suscripción a algunos portales digitales, lo que restringió el acceso a publicaciones completas. Esta limitación podría haber influido en los resultados del análisis de sentimientos, en tanto que la línea editorial de cada medio y sus decisiones sobre qué narrativas destacar o atenuar, pueden modular tanto el contenido temático como el tono emocional de las noticias. En futuras investigaciones sería pertinente analizar las orientaciones

REFERENCIAS

Aho, A. V. (1990). Algorithms for finding patterns in strings. En J. van Leeuwen (Ed.), *Algorithms and Complexity* (pp. 255-300). Elsevier Science Publishers B.V. https://doi.org/10.1016/B978-0-444-88071-0.50010-2

Baker, L. M., & Irani, T. (2014). The impact of new media on policy affecting agriculture. *Journal of Applied Communications, 98*(3), Art. 3. https://doi.org/10.4148/1051-0834.1083

Blei, D. M., Y. Ng, A., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bouchet-Valat, M. (2023). *SnowballC: Snowball stemmers based on the C «libstemmer» UTF-8 Library*. R package version 0.7.1. https://CRAN.R-project.org/package=SnowballC

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*(2), 15-21. https://doi.org/10.1109/MIS.2013.30

Feinerer, I., & Hornik, K. (2024). *_tm: Text Mining Package_*. R package version 0.7-15. https://CRAN.R-project.org/package=tm

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54. https://doi.org/10.18637/jss.v025.i05

Fitzgerald, M. (2012). *Introducing regular expressions: Unraveling regular expressions, step-by-step*. O'Reilly Media, Ed.

Gan, J., & Qi, Y. (2021). Selection of the optimal number of topics for LDA topic model—Taking patent policy analysis as an example. *Entropy*, *23*(10), 1301. https://doi.org/10.3390/E23101301

Grefenstette, G. (1999). Tokenization. En H. van Halteren (Ed.), *Syntactic wordclass tagging. text, speech and language technology*, vol 9 (pp. 117-133). Springer. https://doi.org/10.1007/978-94-015-9273-4_9

Haller, L., Specht, A. R., & Buck, E. B. (2019). Exploring the impact of Ohio agricultural Organizations' social media use on traditional media coverage of agriculture. *Journal of Applied Communications, 103*(4), Art. 4. https://doi.org/10.4148/1051-0834.2264

Happer, C., & Philo, G. (2013). The role of the media in the construction of public belief and social change. *Journal of Social and Political Psychology, 1*(1), 321-336. https://doi.org/10.5964/jspp.v1i1.96

Istat (Istituto Nazionale di Statistica). (2024). *Rapporto annuale 2024 La situazione del Paese*. Istat. https://www.istat.it/produzione-editoriale/rapporto-annuale-2024-la-situazione-del-paese-2/

Jockers, M. (2023). *Extracts sentiment and sentiment-Derived plot arcs from text [R package syuzhet version 1.0.7]*. CRAN: Contributed Packages. https://doi.org/10.32614/CRAN.PACKAGE.SYUZHET

Jockers, M. L. (2015*). Syuzhet: Extract sentiment and plot arcs from text*. GitHub. https://github.com/mjockers/syuzhet

Kr Yadav, M., Kumar Darji, R., & Kumar Yadav, M. (2024). Impact of mass media in agriculture: An overview. *Agricultural and Biological Research, 40*(04), 1232-1235. https://doi.org/10.35248/0970-1907.24.40.1232-1235 / https://www.abrinternationaljournal.org/articles/impact-of-mass-media-in-agriculture-an-overview-110379.html

Liu, B. (2012). Sentiment lexicon generation. En B. Ling (Ed.), *Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies* (pp. 79-89). Springer. https://doi.org/10.1007/978-3-031-02145-9_6

Liu, B. (2017). Many facets of sentiment analysis. En E. Cambria, D. Das, S. Bandyopadhyay, & A. Feraco (Eds.), *A practical guide to sentiment analysis* (pp. 11-39). Springer. https://doi.org/10.1007/978-3-319-55394-8_2

Mohammad, S. M. (2020). *Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text*. arXiv:2005.11882 [cs.CL].

Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a word-emotion association Lexicon.* arXiv:1308.6297 [cs.CL]. http://arxiv.org/abs/1308.6297

Mohr, S., & Höhler, J. (2023). Media coverage of digitalization in agriculture - an analysis of media content. *Technological Forecasting and Social Change, 187,* 122238. https://doi.org/10.1016/j.techfore.2022.122238

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. https://doi.org/10.1561/1500000011

Plutchik, R. (1980). A general psycho evolutionary theory of emotion. En R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience* (pp. 3-33). Elsevier. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29. https://doi.org/10.5120/ijca2018917395

R Core Team. (2022). *R: A language and environment for statistical computing.* R package version 4.2.0. https://www.R-project.org

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE, 16*(8), e0254937. https://doi.org/10.1371/JOURNAL.PONE.0254937

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (1a. ed.). O'reailly. https://www.tidytextmining.com/

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *The Journal of Open Source Software, 1*(3), 37. https://doi.org/10.21105/joss.00037

Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. En *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 88-99). Institute of Formal and Applied Linguistics. http://ufal.mff.cuni.cz/udpipe

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation*. R package version 1.2.1. https://CRAN.R-project.org/package=dplyr