ANÁLISIS DE SENTIMIENTO BASADO EN NOTICIAS

Resultados Preliminares

3 de Junio 2021

EQUIPO DE TRABAJO*

Pilar Cruz N.
Juan Pablo Cova M.
Hugo Peralta V.

CONSTRUCCIÓN DE UN ÍNDICE DE SENTIMIENTO BASADO EN NOTICIAS DE PRENSA

OBJETIVO:

- Complementar data para análisis de actividad económica.

VENTAJAS:

- Datos en tiempo real.
- Alta predictibilidad del ciclo económico.
- Bajos costos en relación a las encuestas.
- Mayor alcance poblacional.
- Alta efectividad en condiciones cambiantes.

OBTENCIÓN DE LOS DATOS

- La base de datos se construyó mediante la recolección de un servicio de noticias.
- A través de técnicas de *webscraping* en Python, se elaboró un robot que permitió la lectura de las noticias y posterior almacenamiento.
- Luego de una limpieza de datos se obtuvieron 420.000 noticias, con más de 100 millones de palabras, en un período comprendido entre enero 2015 – mayo 2020.

LIMPIEZA DE LOS DATOS

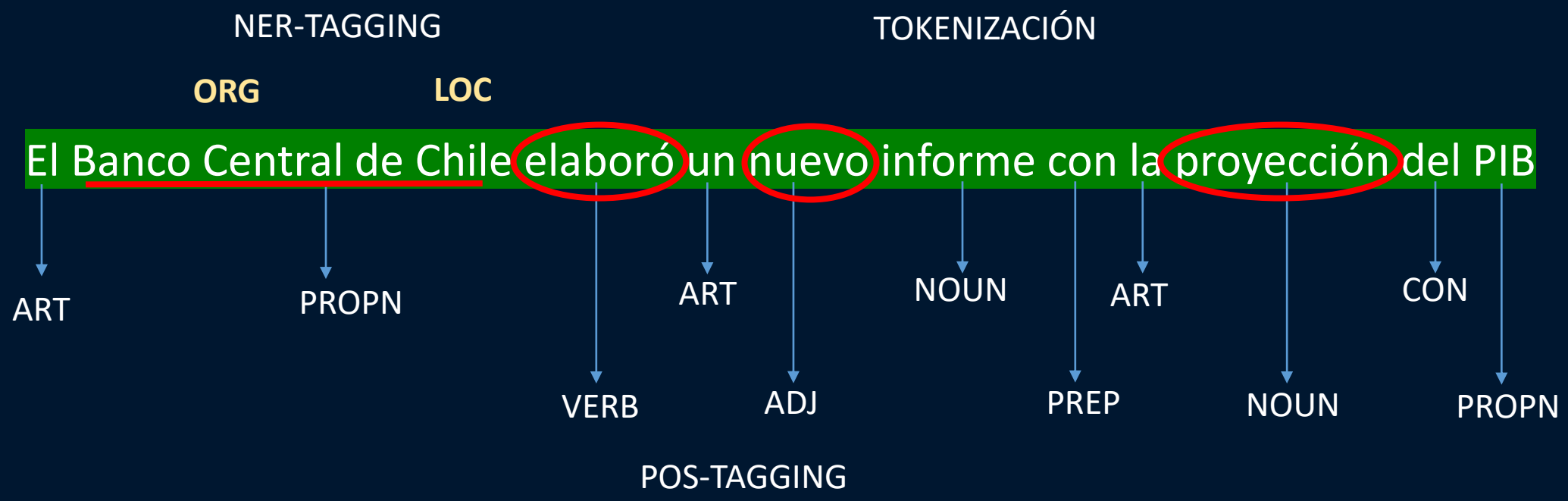
Información de prensa presenta alta complejidad en transformar los datos para su uso en *text mining*.

1. Más de 1000 nombres distintos en los 5 medios utilizados.
2. Descartar noticias muy cortas o demasiado largas.
3. Más 27.000 secciones → Sólo secciones “económicas”
   - Deportes
   - Espectáculos
   - Insertos publicitarios
   - Misceláneos

banco central Chile

PROCESAMIENTO DE LOS DATOS

- Gran cantidad de datos (850K noticias – 1.3 GB texto plano); se debió utilizar formas más eficientes de almacenamiento.
- Se aprovechó la librería “spaCy” que contiene una batería de funciones aplicadas al procesamiento del lenguaje natural (NLP) como son la **lematización**, ***POS-Tagging***, nombramiento de entidades, o la separación por token u oración.



TEXT MINING:

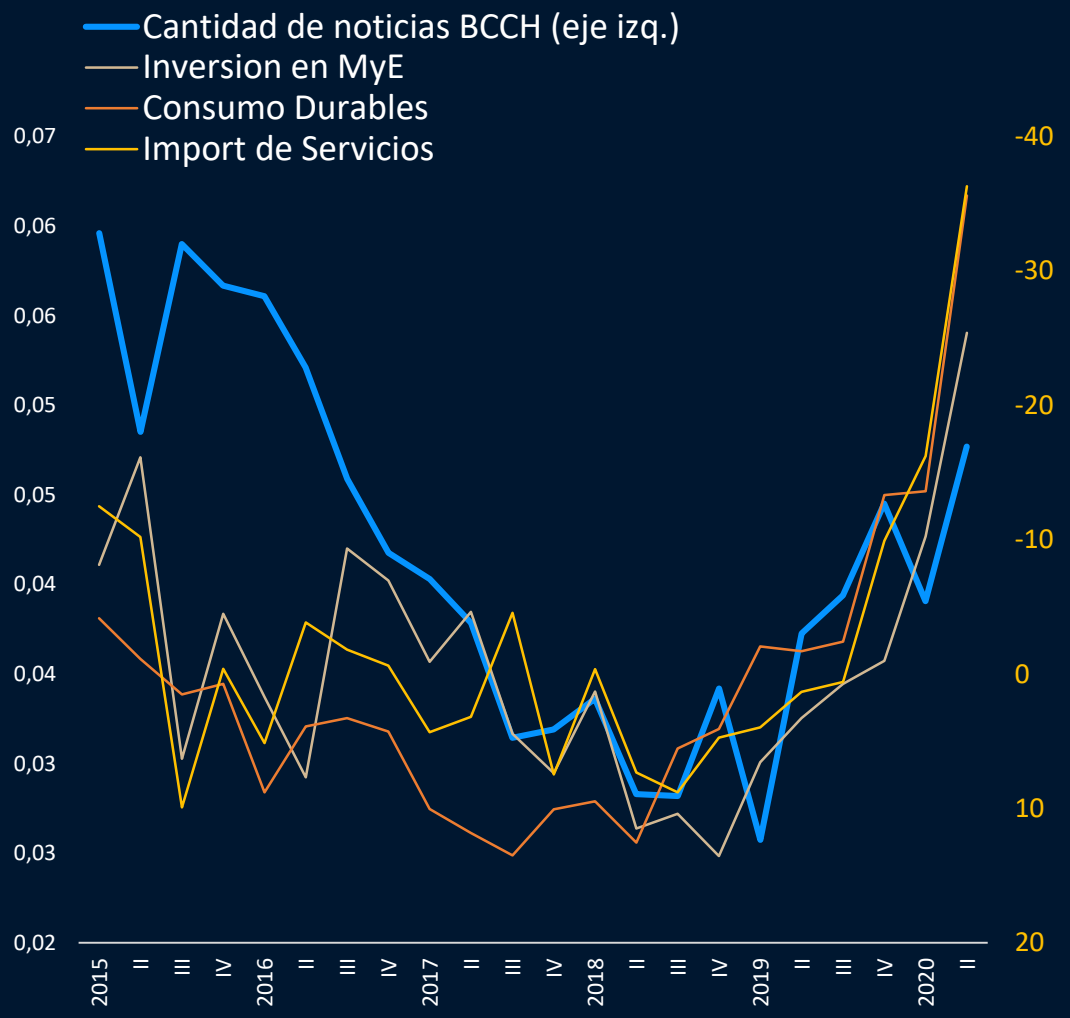
DIVERSIDAD
DE APLICACIONES EN

NOTICIAS

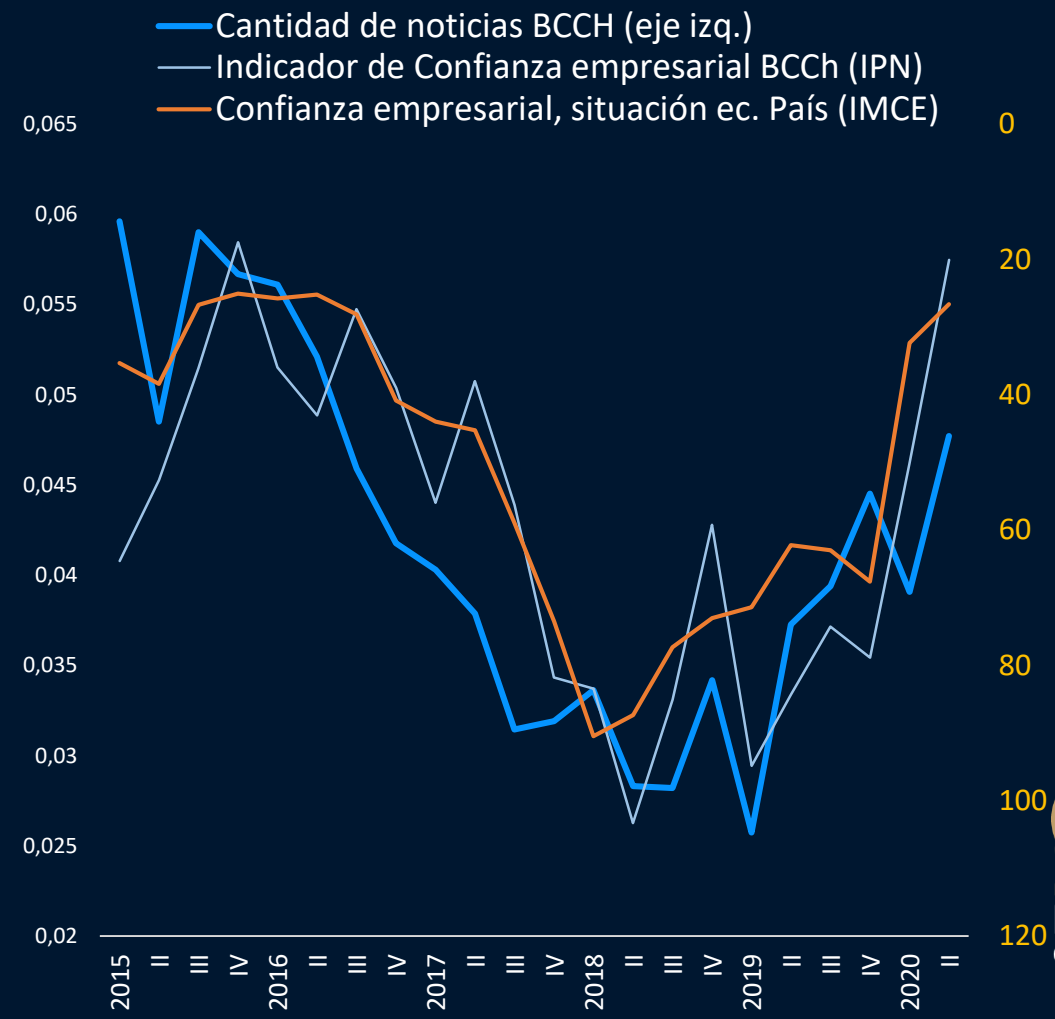
DE PRENSA

MIDIENDO INTENSIDAD: PRESENCIA DEL BCCH EN MEDIOS

SUBE AL CAER INDICADORES DE ACTIVIDAD



SUBE AL CAER INDICADORES DE CONFIANZA



banco central Chile

IS-NEWS

RESULTADOS PRELIMINARES

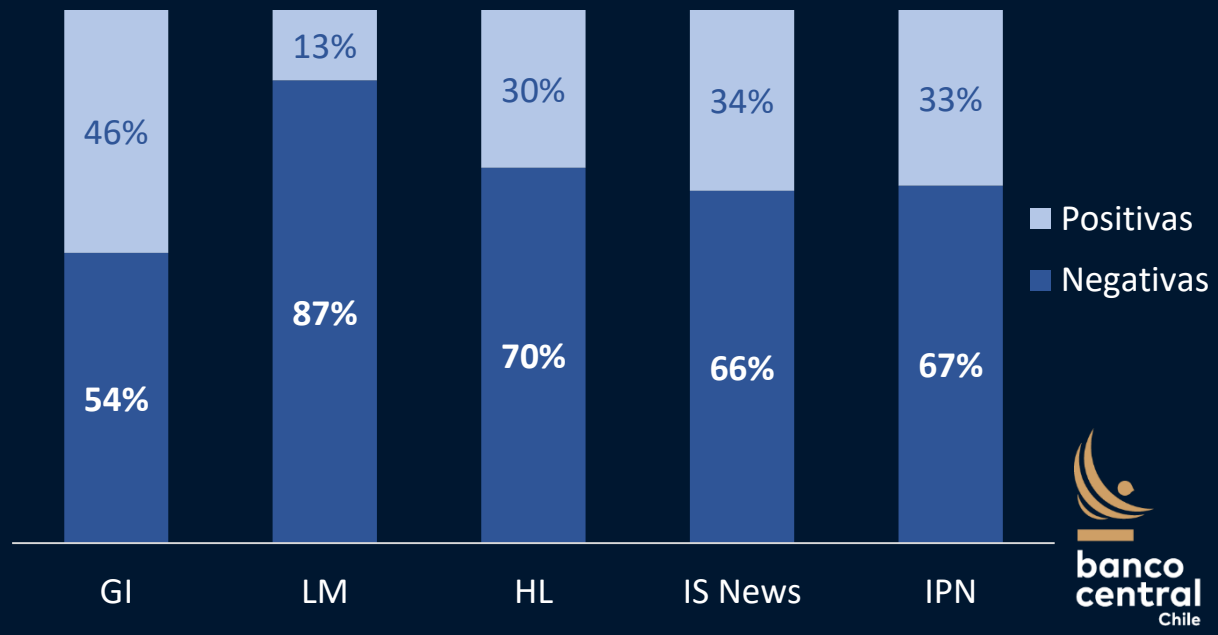
SE CREÓ UN DICCIONARIO EXTENSO BASADO EN NOTICIAS DE PRENSA ESCRITA

- Se identificaron verbos, adjetivos y adverbios presentes en las noticias.
- Se revisaron 2.832 verbos, 4.157 adjetivos y 627 adverbios = 7.616 palabras únicas (10% de su total en la lengua española).
- Se etiquetaron distinto de cero, 257 verbos, 67 adjetivos y 50 adverbios únicos (374 en total), equivalentes a 5.619 palabras en sus diversas formas flexionadas (conjugación y número).

TAMAÑO DE PRINCIPALES DICCIONARIOS ETIQUETADOS

Diccionario	Especialidad	Idioma	Total palabras etiquetadas
Harvard General Inquirer (GI)	Inglés General	Inglés	4.206
Loughran-McDonald (LM)	Estados Financieros Empresas (financiero)	Inglés	2.683
Hu-Lui (HL)	Reseña de Películas	Inglés	6.789
IS News	Prensa Escrita Chilena	Español	5.619
IPN BCCh	IPN BCCh	Español	678

PALABRAS CON ETIQUETAS POSITIVAS Y NEGATIVAS EN DICCIONARIOS INTERNACIONALES (% del total)



banco central Chile

TAMBIÉN SE ETIQUETARON MANUALMENTE NOTICIAS

PARA CALIBRAR EL DICCIONARIO

Se seleccionaron 840 noticias: [1]

- Diarios más relevantes.
- Palabras por noticia: entre 200 y 400.
- Incluyen : “mencionó”; “explicó”; “dijo”; “escribió”; “señaló”, para captar tono.
- Selección aleatoria

Clasificación manual de los textos: 23 personas del BCCH[2].

[1] Fuente: *Shapiro et al., 2020*.

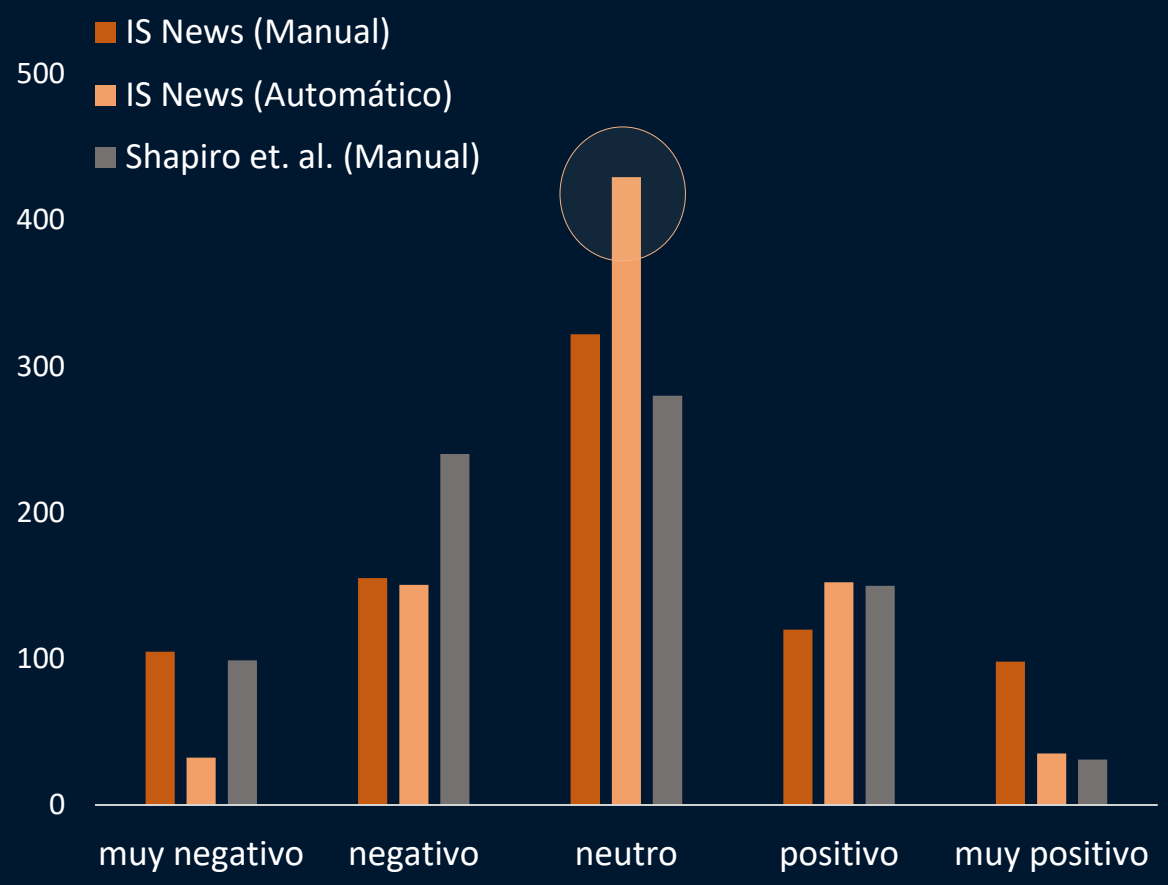
[2] Se utilizo la plataforma digital de *e-learning* del *startup Boots.*

banco central Chile

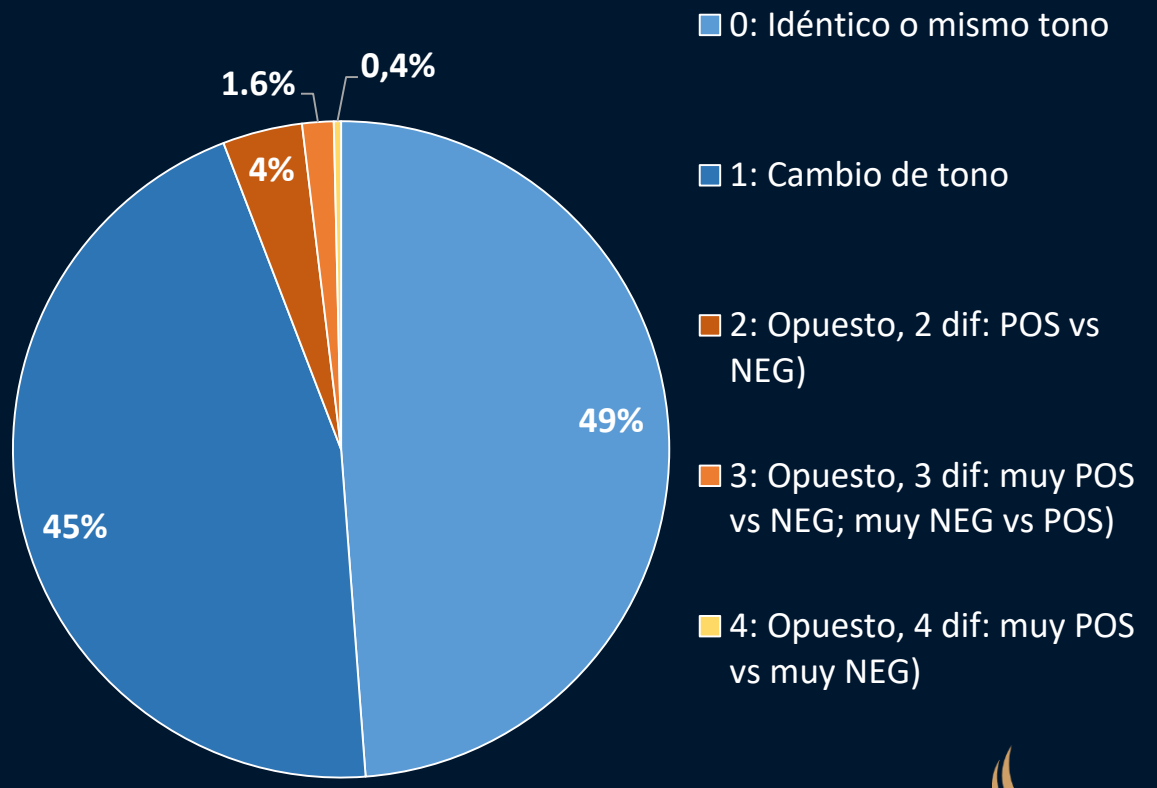
NOTICIAS: ETIQUETADO AUTOMÁTICO vs MANUAL

Similares aunque etiquetado automático presenta menor varianza y sesgo hacia noticias neutrales

HISTOGRAMA IS-NEWS: COMPARATIVO ETIQUETADO MANUAL VS AUTOMÁTICO



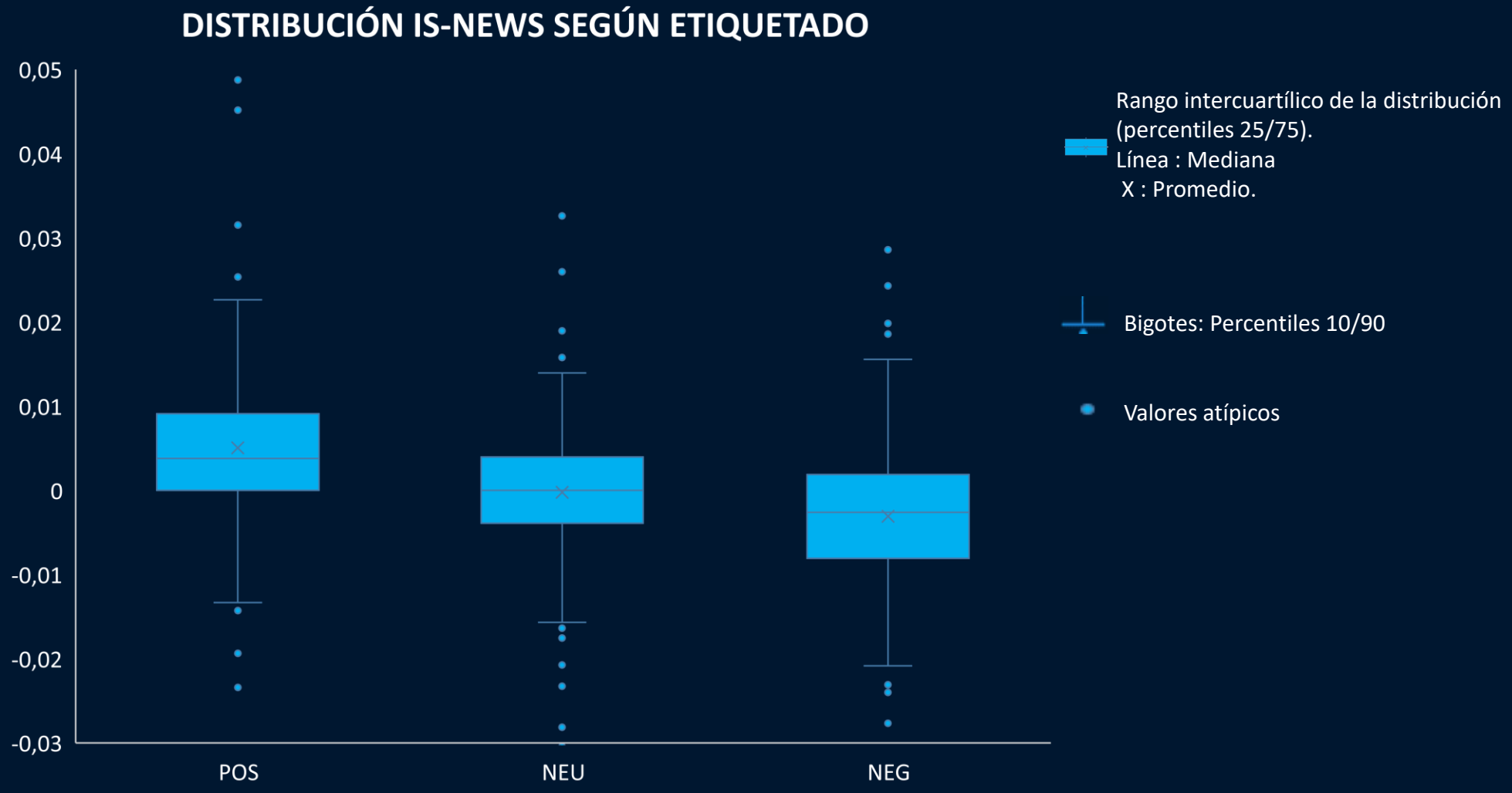
RESULTADOS: ETIQUETADO MANUAL VS AUTOMÁTICO



banco central Chile

NOTICIAS: ETIQUETADO AUTOMÁTICO vs MANUAL

Distribución etiquetado automático consistente con manual, aunque con algunos valores atípicos.



banco central Chile

CONSTRUCCIÓN DE UN PRIMER INDICADOR

- Con un diccionario etiquetado en las categorías de verbos, adjetivos y adverbios, como sentimiento o modificadores, se calcula un indicador simple.
- La regla para el cálculo se puede describir bajo el siguiente algoritmo:

Para cada noticia de la base de datos:

Luego para cada oración de la noticia:

Luego para cada token (palabra) de la oración:

Si pertenece a la categoría de VERBO, ADJETIVO o ADVERBIO:

1. **Si tiene un valor de sentimiento → 1 o -1**
2. si tiene un valor de modificador → 0,5 o 1,5
3. **En cualquier otro caso → 0**

Pasar a la siguiente oración

Al final de la oración, sumar todos los valores de sentimiento y multiplicar el resultado por cada valor modificador

Normalizar indicador por largo del texto y pasar a la siguiente noticia

Ejemplo con 2 oraciones

Sentimiento O1: 4
Modificadores O1: 1,5; 1,5; 0,5
Valor 01: 4*1,5*1,5*0,5 = **4,5**

Sentimiento O2: 0
Modificadores O2: 1,5; 1,5; 0,5
Valor 02: **0**

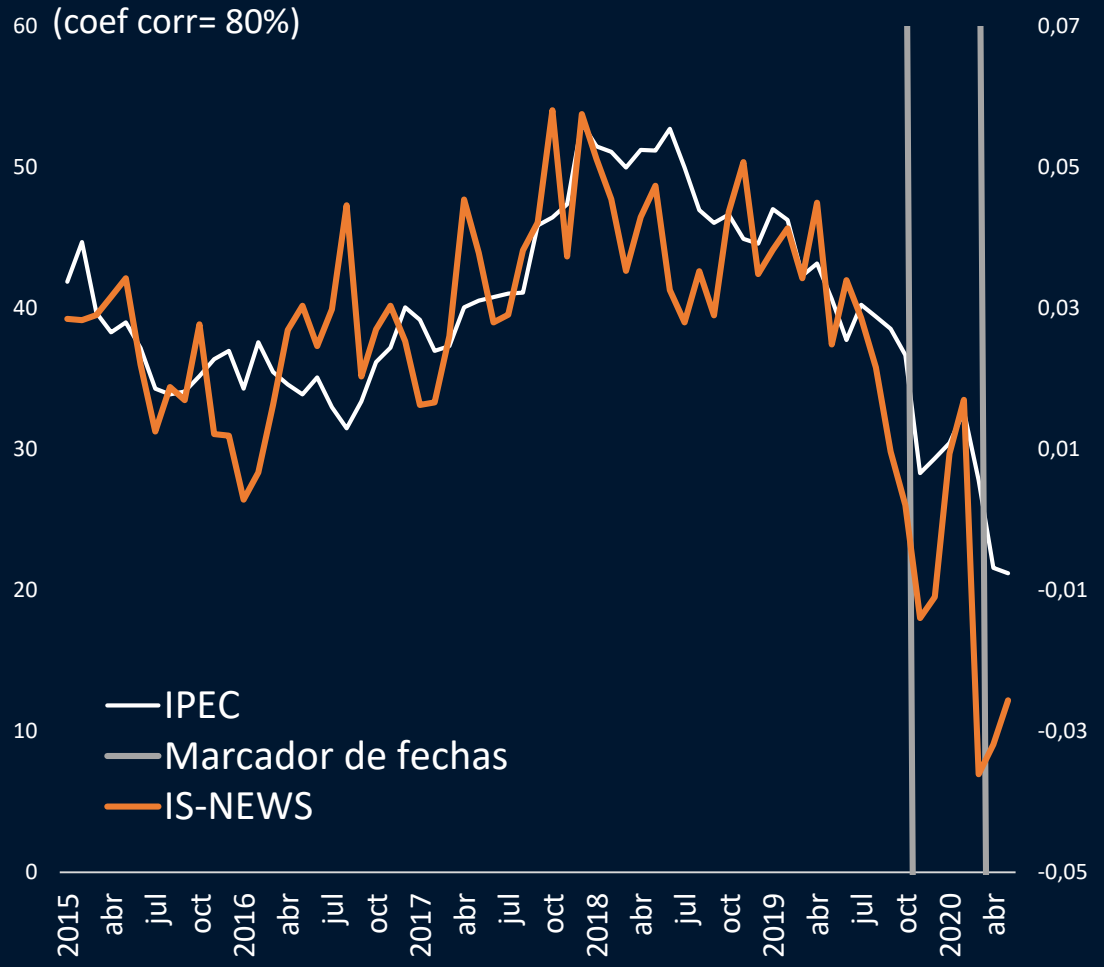
Largo: 120

VALOR FINAL= **0,0375**

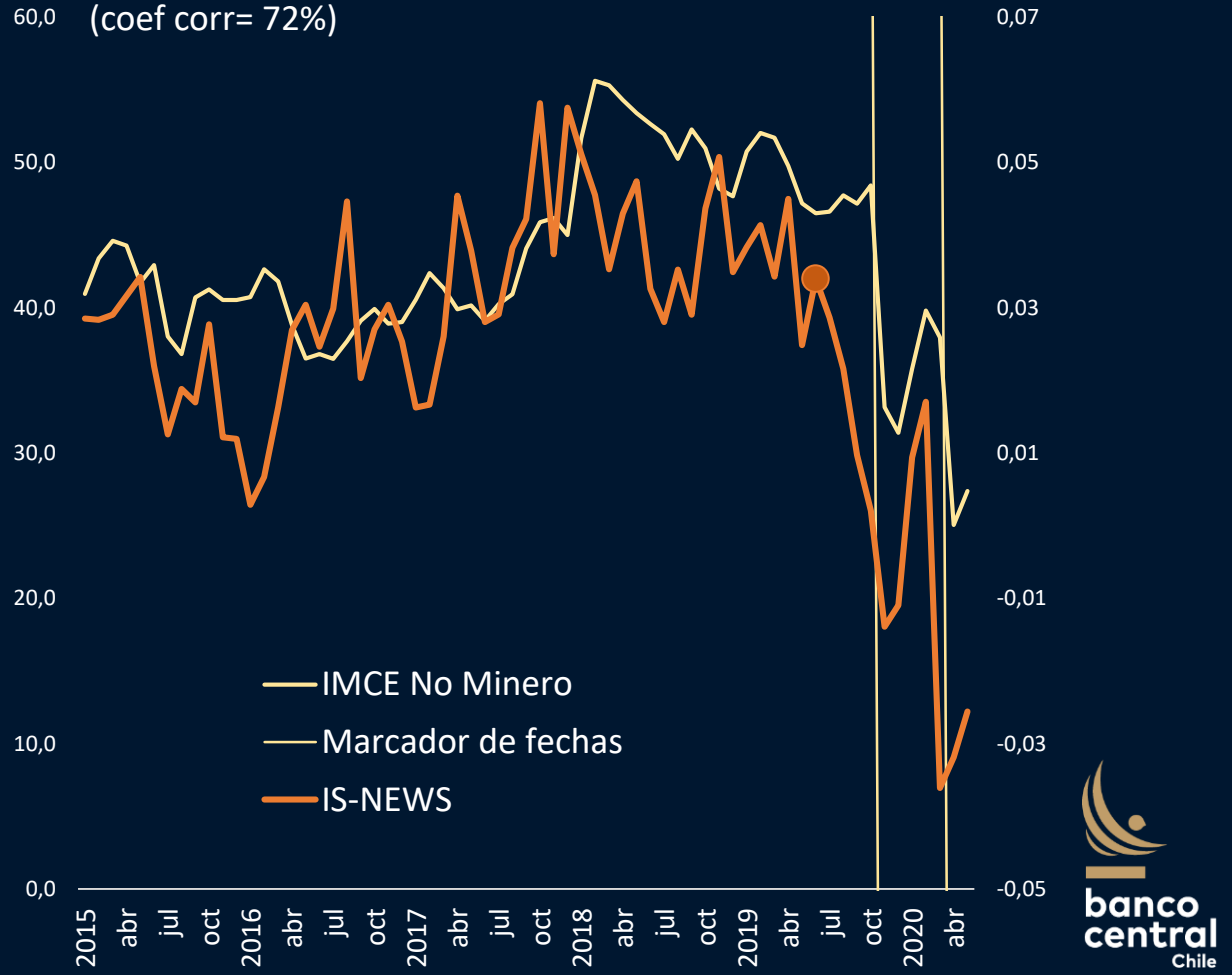
banco central Chile

IS-NEWS SE CORRELACIONA CON INDICADORES DE CONFIANZA

IS - NEWS VS INDICADOR DE CONFIANZA DEL CONSUMIDOR (IPEC) 2015-20



IS - NEWS VS CONFIANZA EMPRESARIAL EXCL. MINERÍA (IMCE)

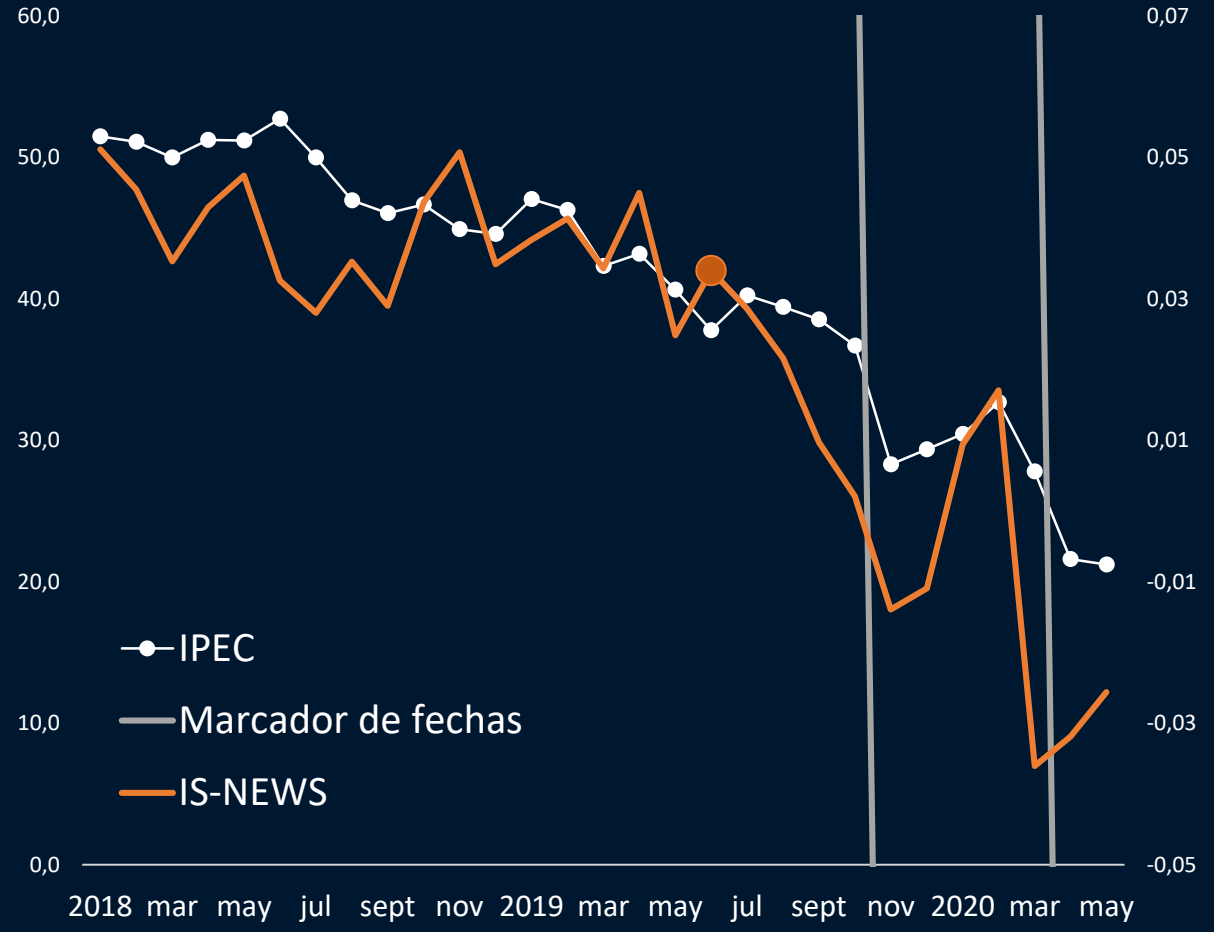


banco central Chile

INDICES DE CONFIANZA: UN ZOOM A LA CRISIS DE OCTUBRE 2019 Y MARZO 2020

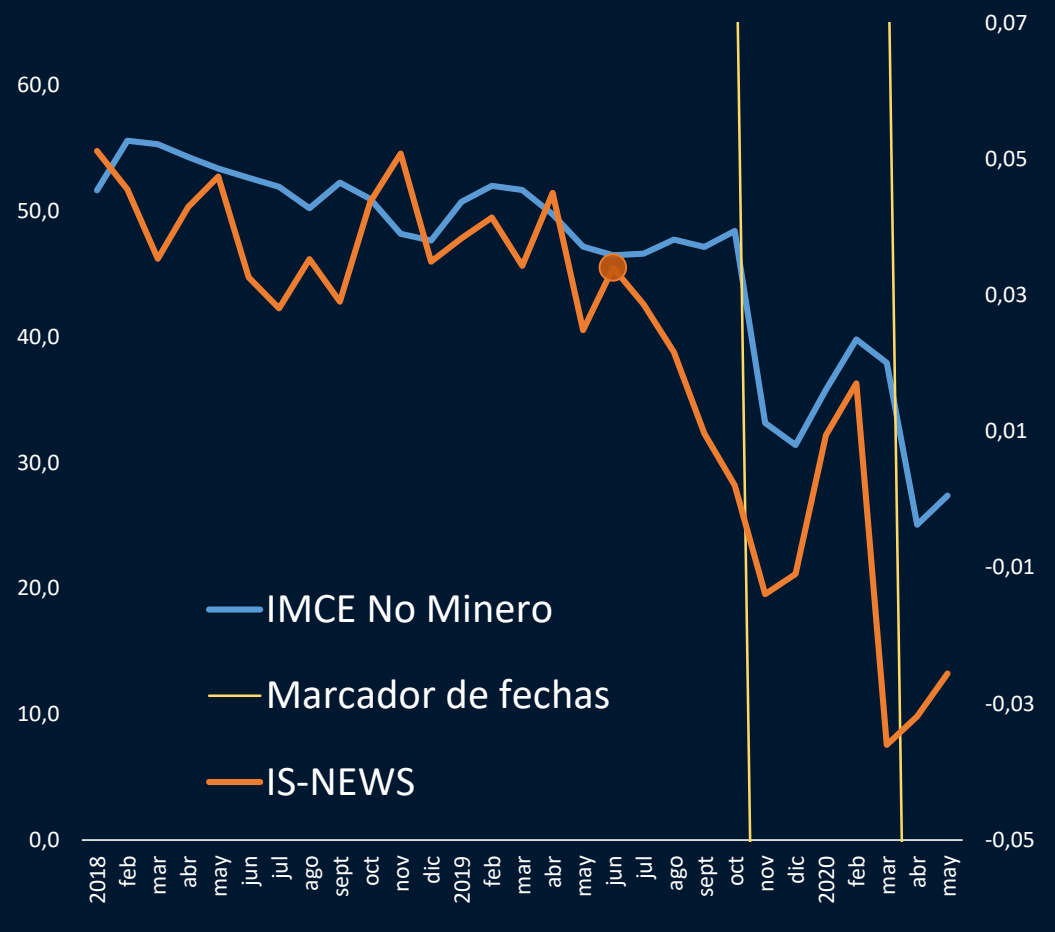
IS - NEWS VS INDICADOR DE CONFIANZA DEL CONSUMIDOR (IPEC) 2018-21

(coef corr= 80%)



IS - NEWS VS CONFIANZA EMPRESARIAL EXCL. MINERÍA (IMCE)

(coef corr= 72%)

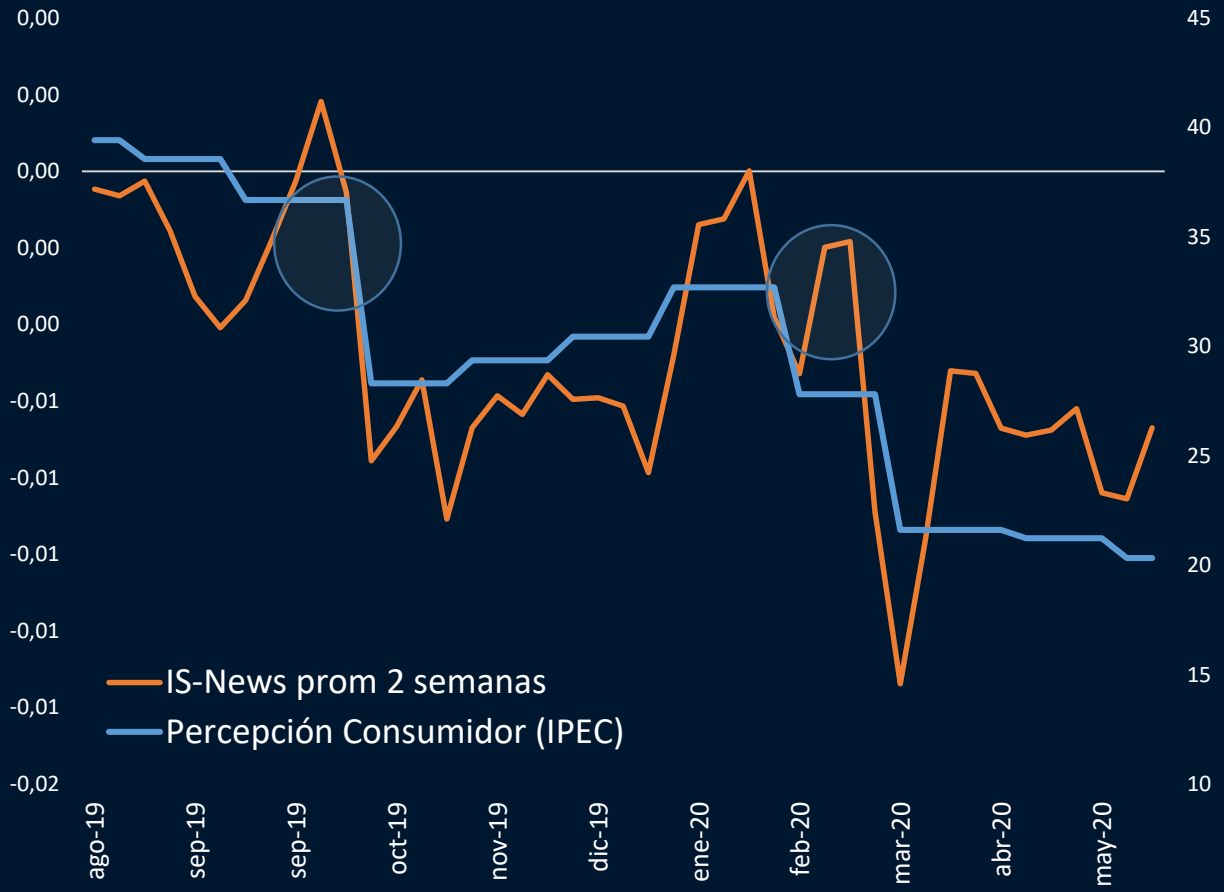


banco central Chile

UN ZOOM SEMANAL

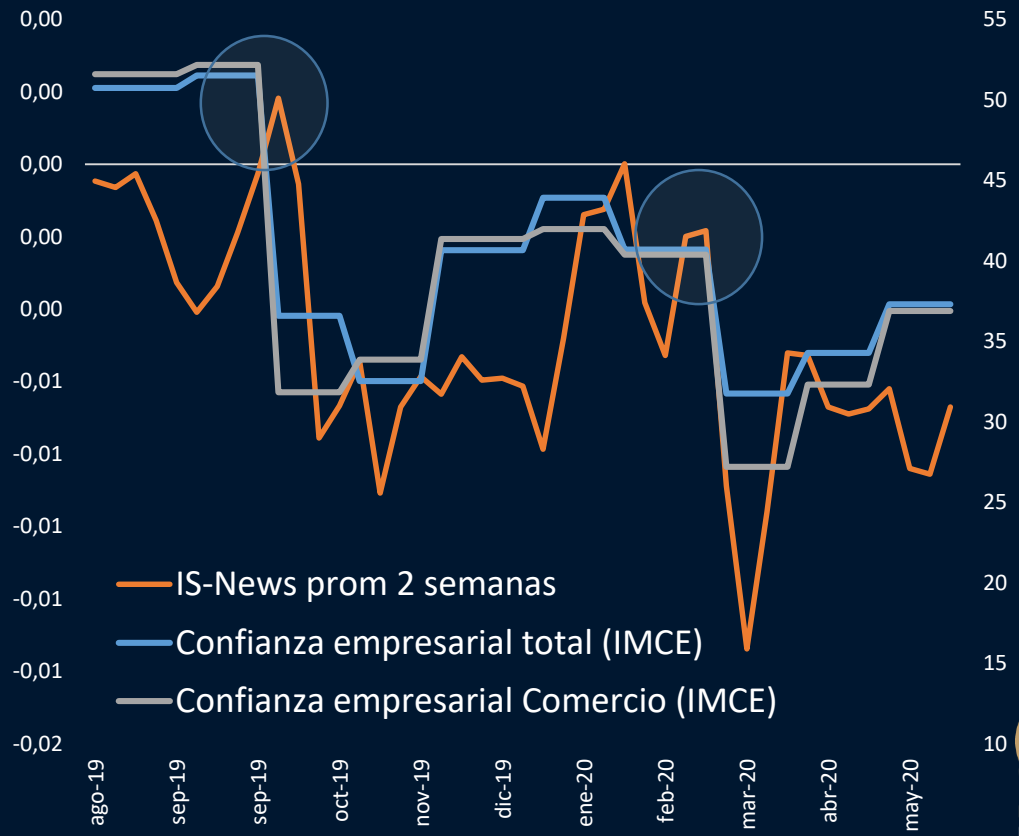
IS-NEWS SEMANAL: ANTICIPA 3 SEM CRISIS SOCIAL Y PANDEMIA- Perc. consumidores

(var. 12 meses, %, prom. móvil 2 sem., desfasado 3 sem.)



IS-NEWS SEMANAL ANTICIPA 4 SEM LA PANDEMIA-Confianza empresarial

(var. 12 ms, %, prom. móvil 2 sem., desfasado 4 sem.)

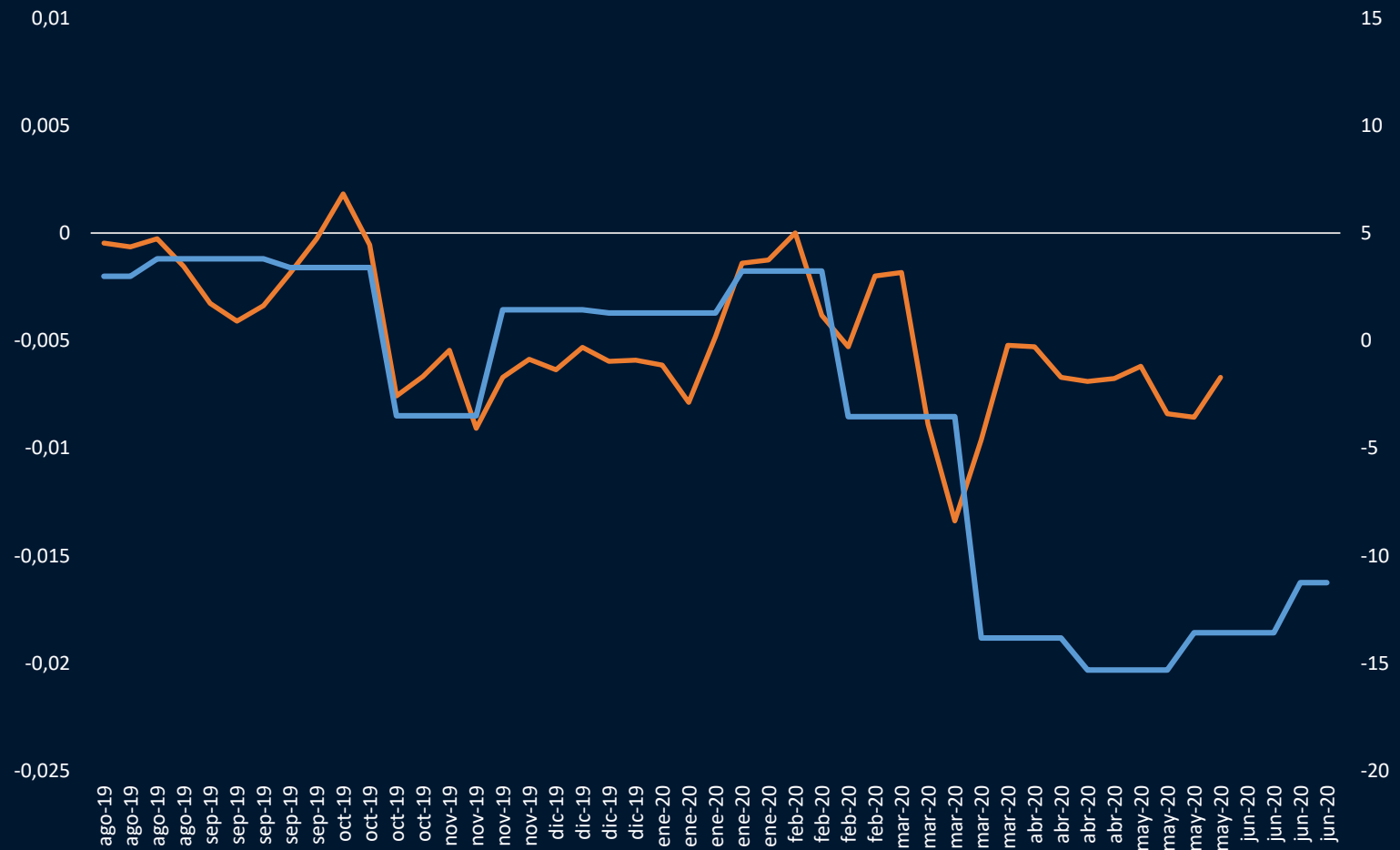


banco central Chile

CIFRAS SEMANALES: PREDICTIBILIDAD DEL CRECIMIENTO ECONÓMICO

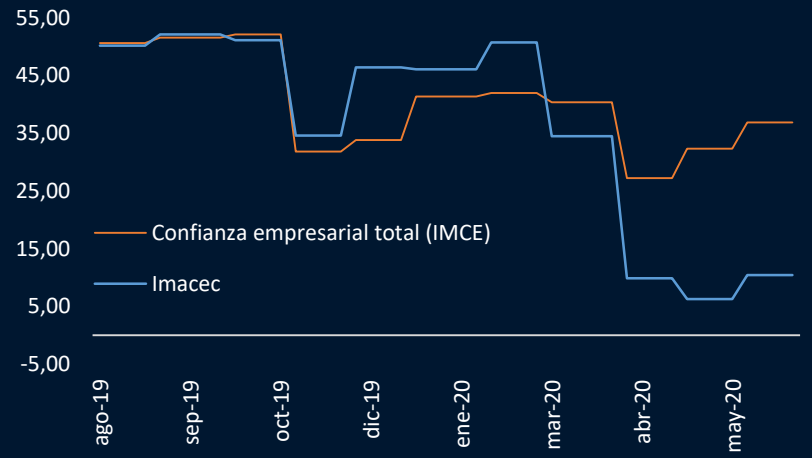
IS-NEWS SEMANAL VS CRECIMIENTO ECONÓMICO (IMACEC)

(variación 12 meses, %, promedio móvil 2 semanas, desfase 2 semanas)



CONFIANZA EMPRESARIAL VS IMACEC

(variación 12 meses, %, promedio móvil 2 semanas)

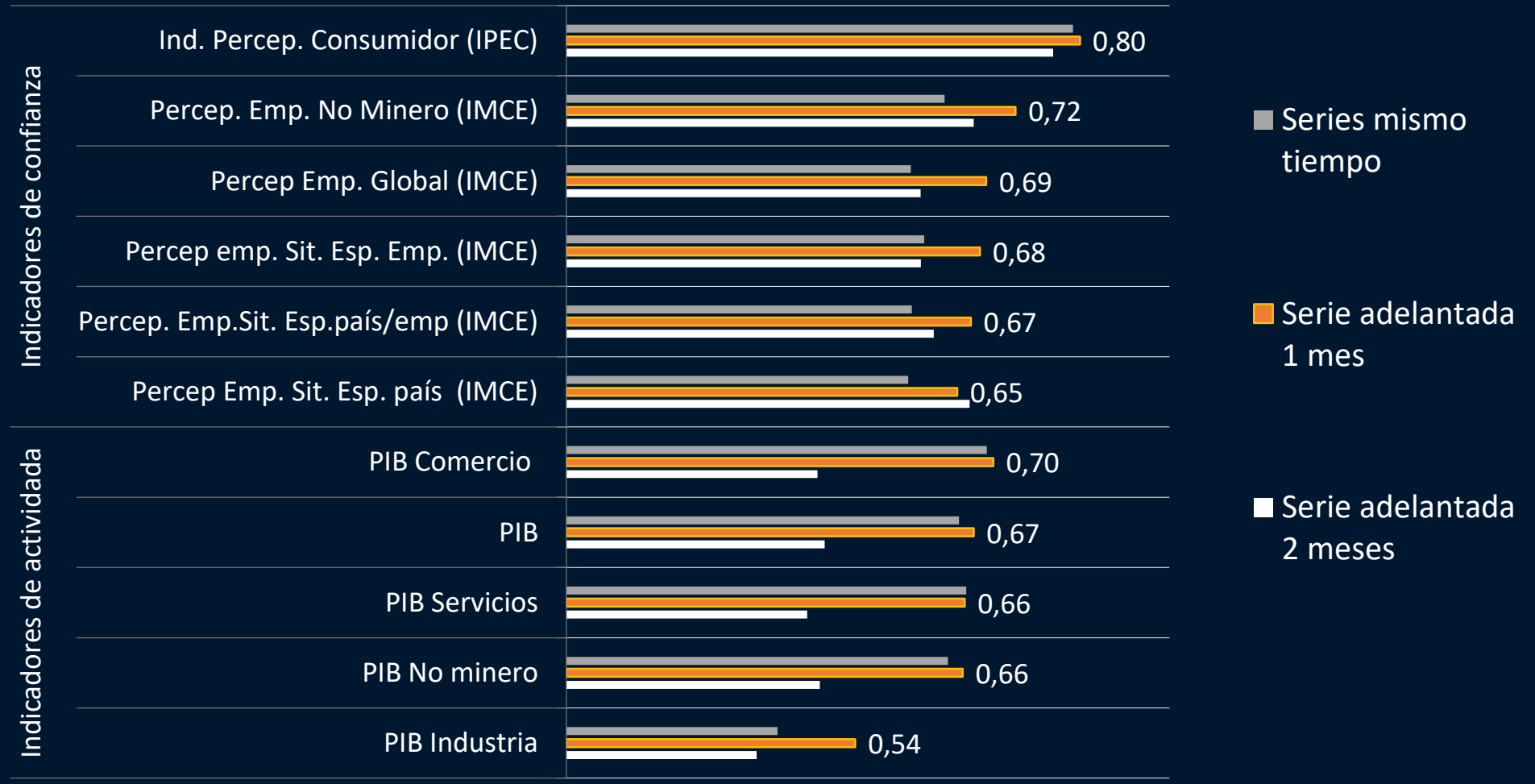


banco central Chile

IS-NEWS SE CORRELACIONA CON INDICADORES DE CONFIANZA Y COMERCIO

IS – NEWS: INDICADOR LÍDER ACTIVIDAD Y CONFIANZA

Índices de confianza en bases; índices de actividad: var 12 meses, %



banco central Chile

COMENTARIOS FINALES

1. **IS-News**: altas correlaciones con índices de confianza y de actividad.

2. **Efectividad:** basada en la construcción de un diccionario de propósito, y dimensión similar a los de mayor uso en el idioma inglés.

3. **Predictibilidad:** anticipa los shocks económicos en la economía chilena en un lapso de alrededor de 3 a 4 semanas.

4. **Alta disponibilidad y bajo costo:** la implementación de un indicador de prensa en tiempo real e independiente a otras fuentes de datos.

5. ***Text mining* en noticias de prensa**: numerosas aplicaciones adicionales, como es el análisis de tópicos y bolsas de palabras para medir intensidad.

6. **Desarrollos complementarios al IS-News**: IS con modelos de *machine learning* en español (*BERT*), o tipo *Vader* en inglés, en lugar de léxicos etiquetados.

banco central Chile

ANÁLISIS DE SENTIMIENTO BASADO EN NOTICIAS

Resultados Preliminares

3 de Junio 2021

EQUIPO DE TRABAJO*

Pilar Cruz N.
Juan Pablo Cova M.
Hugo Peralta V.