BANCO DE MÉXICO®

Propuesta de un Índice de riesgo y de métodos de clasificación de textos, basados en el reporte de Actividad Económica Regional.

Erika Tatiana Rueda Santos

Victor Manuel Gómez Espinosa

Tutores:

Dr. Víctor Hugo Muñiz Sánchez

Dr. Jorge Alberto Alvarado Ruiz

CONACYT Consejo Nacional de Ciencia y Tecnología

CIMAT Centro de Investigación en Matemáticas, A.C.

MTY Modelando el siglo XXI

Propuesta de índice de riesgo

Problema

•El **Banco de México** entre sus diferentes actividades, **por medio de encuestas a empresarios** (**y expertos** en el tema), **recopila de manera trimestral información textual a nivel regional** acerca de **sus expectativas sobre la economía mexicana**.

•Ante esto, se tiene como **objetivo principal** la creación de **indicadores que reflejen el impulso o la limitación de la actividad económica** tanto **del país** como **a nivel regional**, basados en datos de texto.

Problema

•Este problema **ya ha sido analizado anteriormente**, extrayendo el sentimiento de cada texto **mediante un Lexicón** para posteriormente crear los indicadores (Miranda 2020), sin embargo, el enfoque que se pretende **en el trabajo actual** es diferente, ya que **consiste en la identificación de tópicos relevantes por métodos no supervisados de aprendizaje máquina** para la generación de los índices.

Métodos para clusters de tópicos

Para encontrar clusters de tópicos, entendiéndolo como grupos de palabras semánticamente relacionadas, hay varios métodos en la literatura, como lo son: **métodos probabilísticos** (LDA), métodos que utilizan bolsa de palabras (BOW o TF-IDF) llamados **métodos de factorización de matrices**, entre otros

Métodos para clusters de tópicos

En este trabajo se emplea un método diferente a los anteriores, que consiste en la **obtención de la representación vectorial de los textos** (embeddings) usando **FastText** ya que tiene la **ventaja** de que se pueden obtener las **representaciones vectoriales de palabras fuera del vocabulario** y posteriormente **clusterizar asociaciones semánticas mediante Fuzzy KMeans**.

Adicionalmente, **la metodología propuesta** en este trabajo tiene la **ventaja** de que **permite rápidamente obtener resultados para todo el periodo de tiempo** con el que se tienen datos (1T11-4T20) **y también para nuevos textos** que se agreguen en el futuro.

Datos

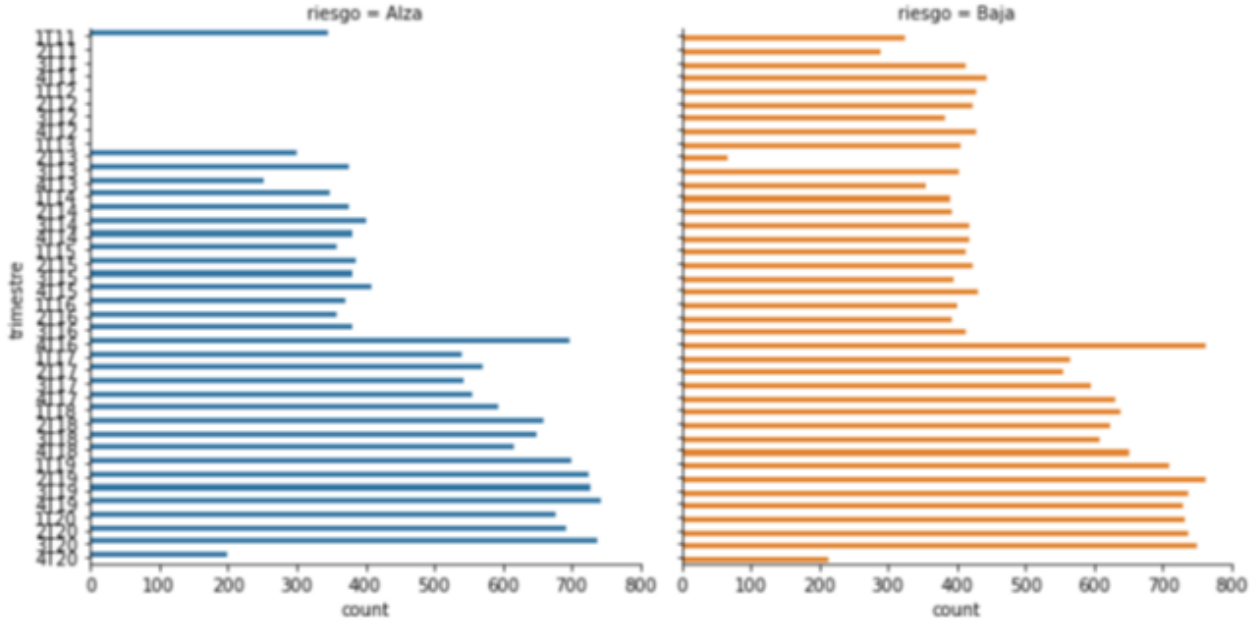


Figura 2.1: Distribución de documentos por trimestre de 2011-2020, según el tipo de riesgo.

Datos

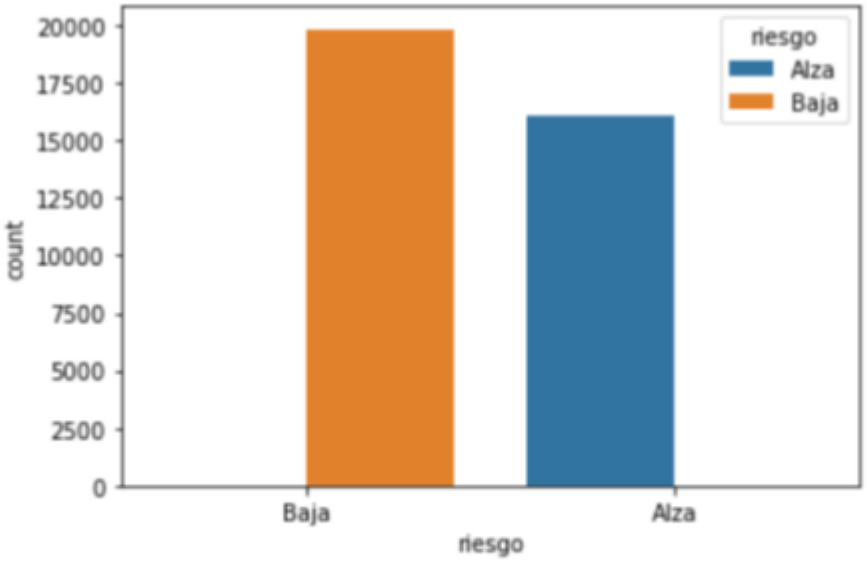


Figura 2.2: Documentos por tipo de riesgo (Baja 55%, Alza 45%). Documentos totales: 35,895.

Datos

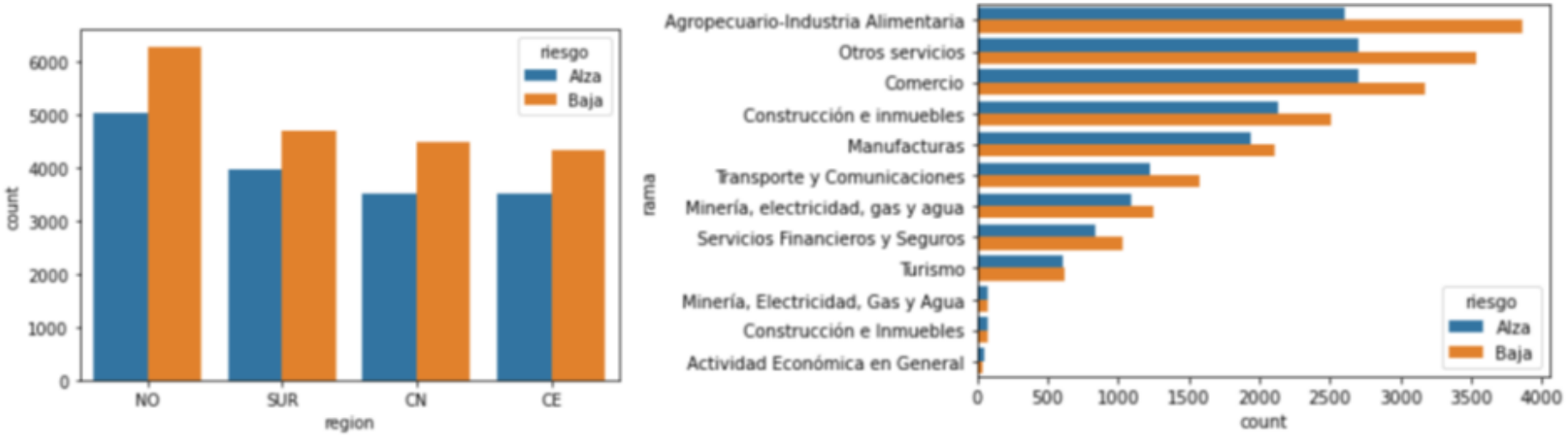


Figura 2.3: a) Documentos por región (NO 31%, CE 22%). b) Documentos por rama (Agropecuario 18%, Actividad económica general 0.2%).

Datos

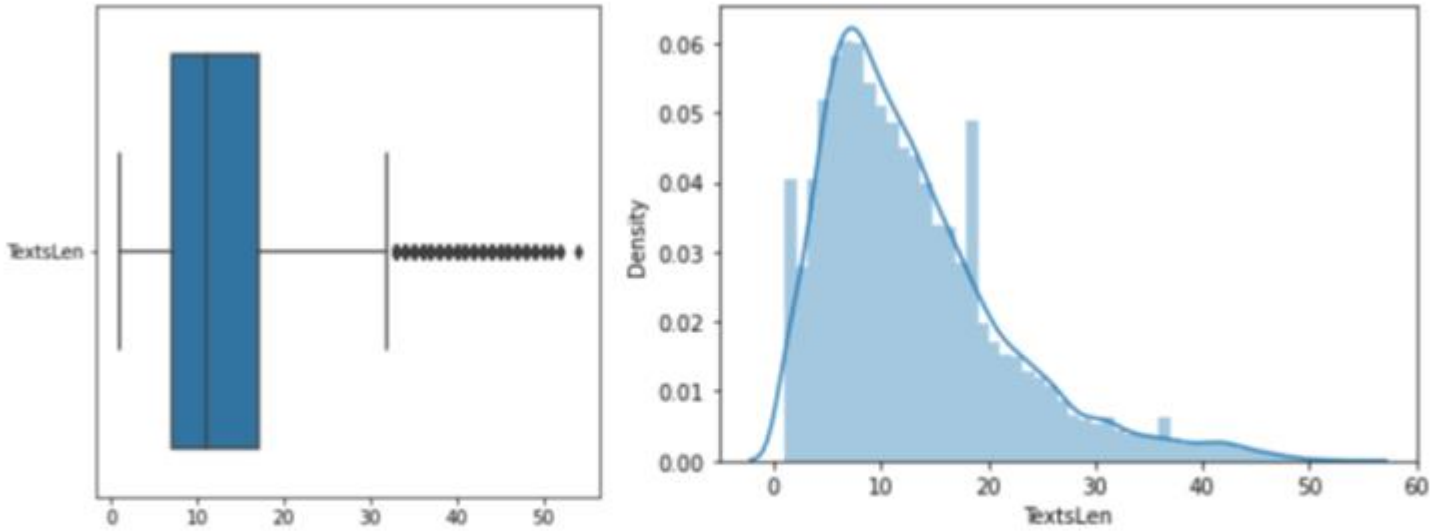


Figura 2.4: Distribución de la longitud de los documentos. Min: 1, 25%: 7, 50%: 11, media: 13, 75%: 17, Max: 54.

Conjuntos de tópicos

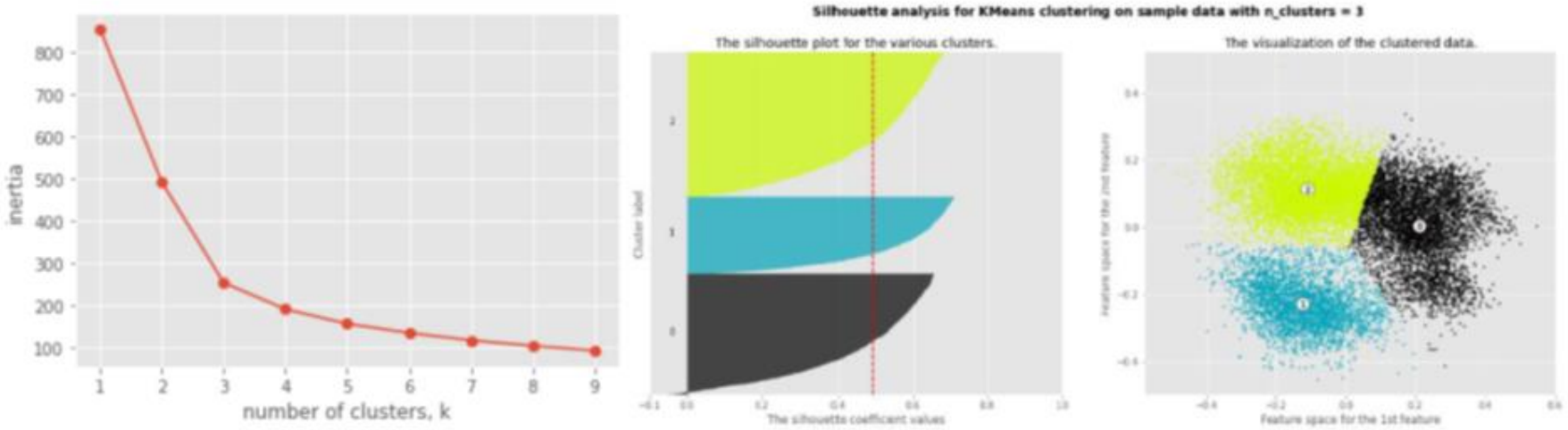


Figura 2.5: a) inercia vs Clústers, punto de cambio en 3. b) Gráfico de silueta, mejor puntuación para 3 clústers: 0.4921.

Conjuntos de tópicos

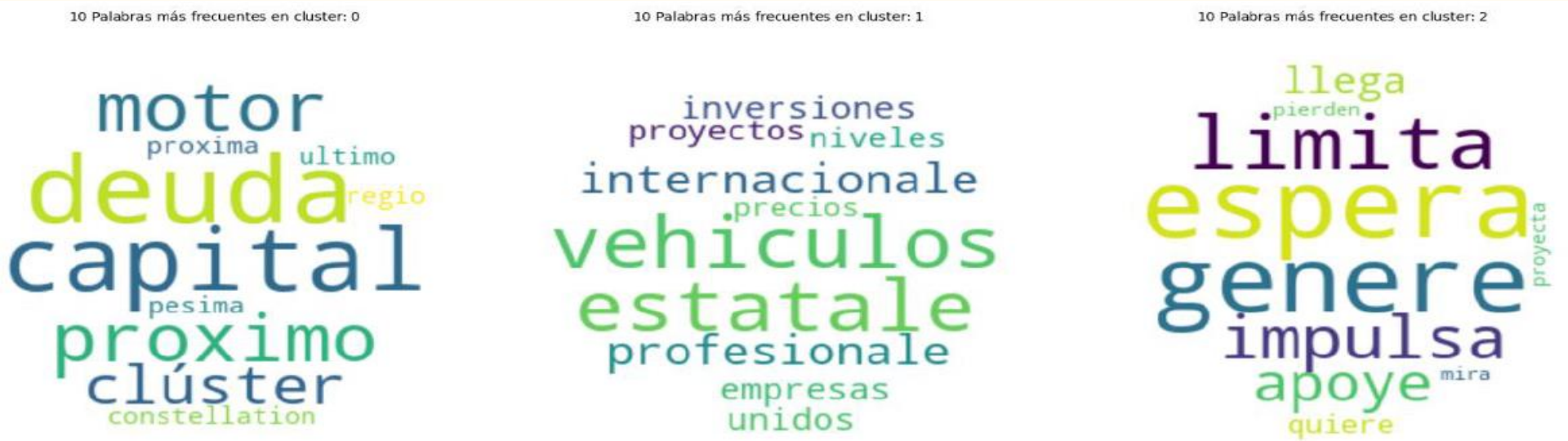


Figura 2.7: Gráficos de nubes de palabras para los clústers: 0, 1, 2.

Conjuntos de tópicos

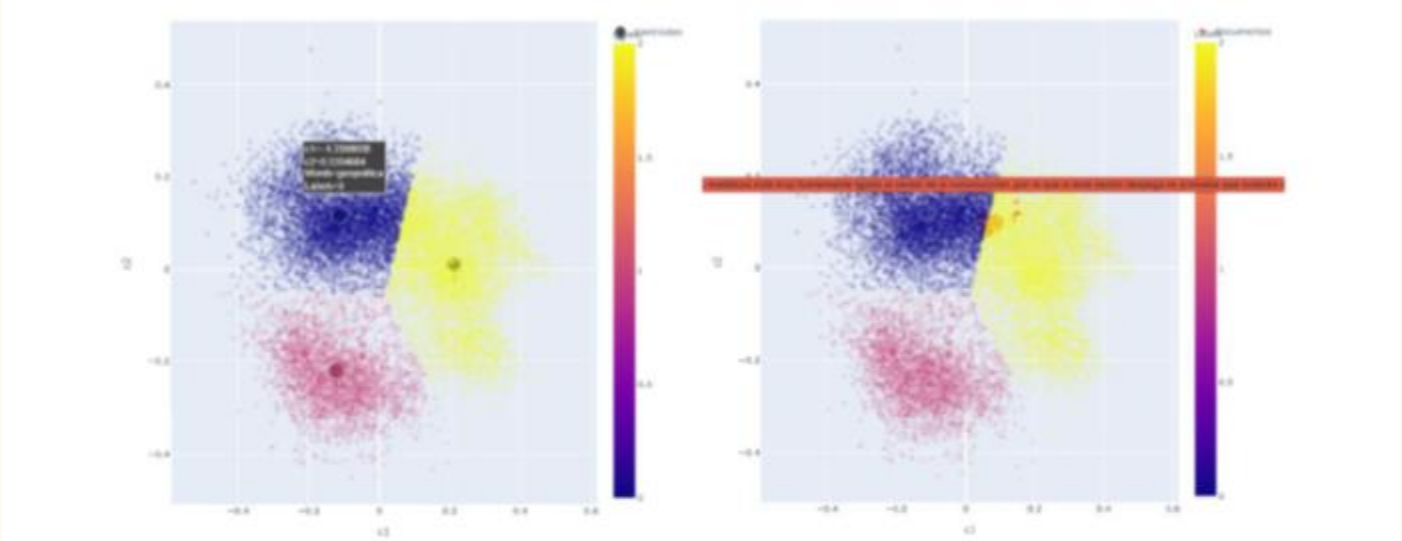
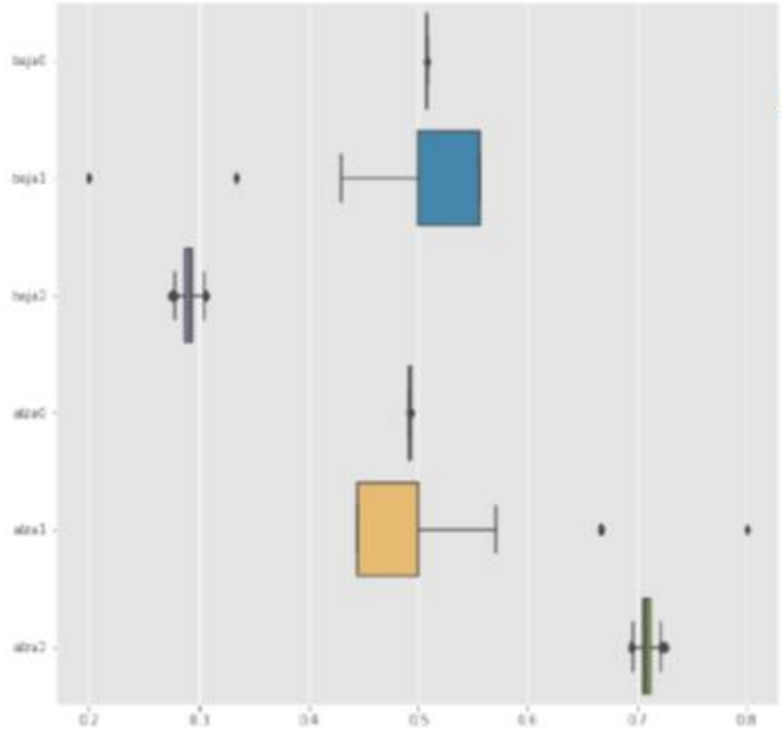


Figura 2.6: a) Clústers de tópicos únicos en el vocabulario (14,475) con sus respectivos centroides, Técnica de clustering empelada: Fuzzy KMeans. b) proyección de los documentos de un trimestre sobre los clústers de tópicos, se asigna el cluster del centroide más cercano al texto.

Conjuntos de tópicos



$$P(B/C_i) = I_{b,i} / N_i$$

Figura 2.8: Frecuencias relativas de los riesgos a la baja o al alza por cada clúster (0,1,2).

Indicadores

$$P(C_i)_Q = {}^{I_{i,Q}}\!/_{N_Q} \tag{1.2}$$

Donde, $\left(P(C_i)_Q\right)$ es la frecuencia relativa para el clúster $(i=0,1,2)$ en el trimestre (Q), $\left(N_Q\right)$es el número de textos en el trimestre y $\left(I_{i,Q}\right)$ es el número de textos clasificados en el clúster en el trimestre.

$$Risk\_index = P(B)_Q = \sum_{i=0}^{2} P(C_i)_Q \, P(B/C_i) \tag{1.3}$$

Donde $P(B)_Q$ es la probabilidad de que para ese trimestre sea riesgo a la baja.

Indicador, General

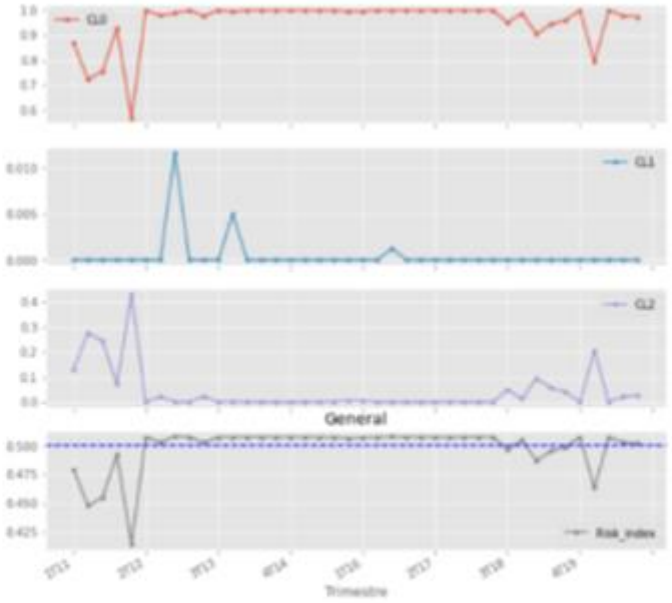


Figura 3.2: Indicador de riesgo con sus respectivas componentes para el periodo de 1T11 a 4T20.

Indicador, General

Figura 3.1: Gráficos de nubes de palabras, con las 50 palabras más representativas en el cluster 0, para el trimestre a) 3T12, b) 3T18.

Indicador, General

Figura 3.3: Gráfico de nube de palabras, para las 50 palabras más frecuentes en el a) clúster 0 y b) clúster 2 para el trimestre 1T20.

Indicador, General

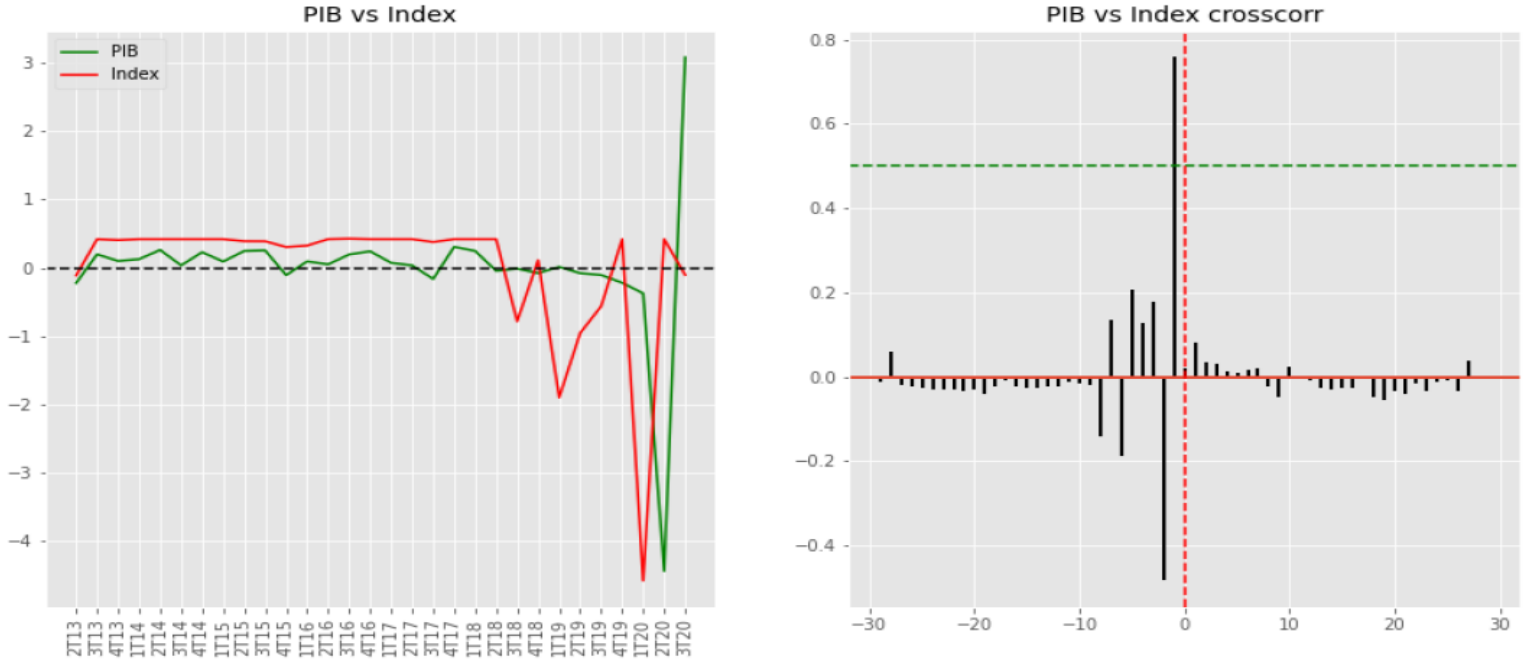


Figura 3.4: a) Indicador general contra el **PIB** del 2T13 al 3T20, b) correlación cruzada para el indicador y el **PIB**.

Indicador, Regional

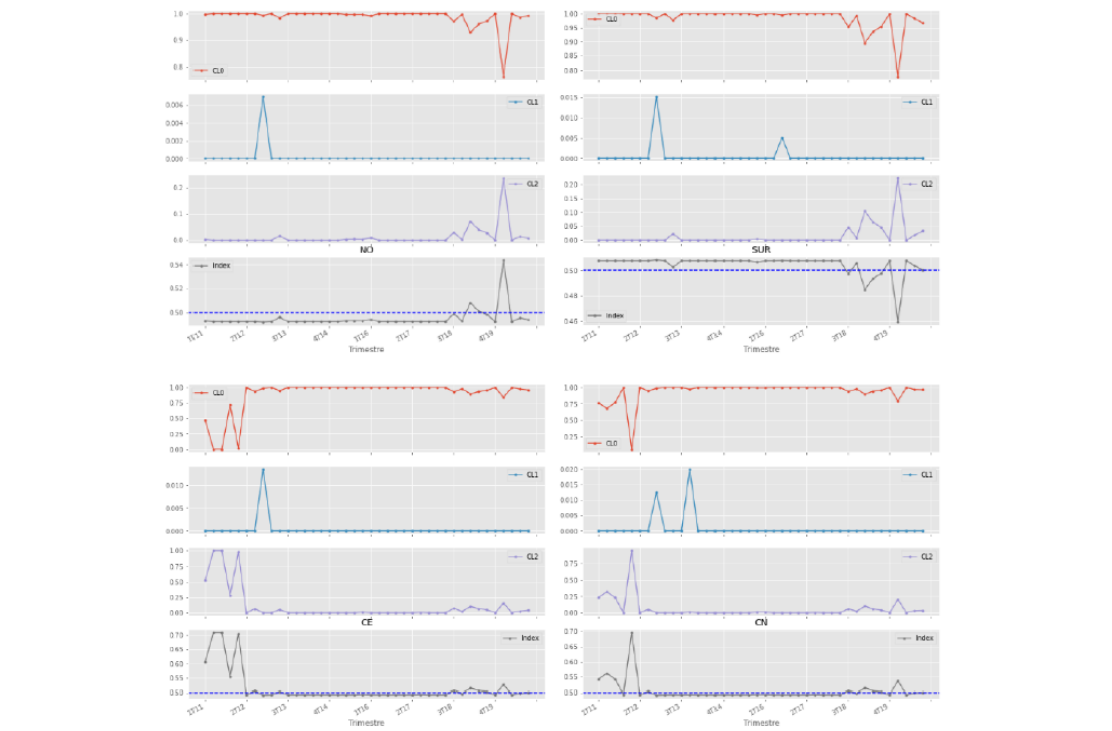


Figura 3.5: Indicadores regionales con sus respectivas componentes para el periodo de 1T11 a 4T20, para todas las regiones.

Indicador, Regional: CE

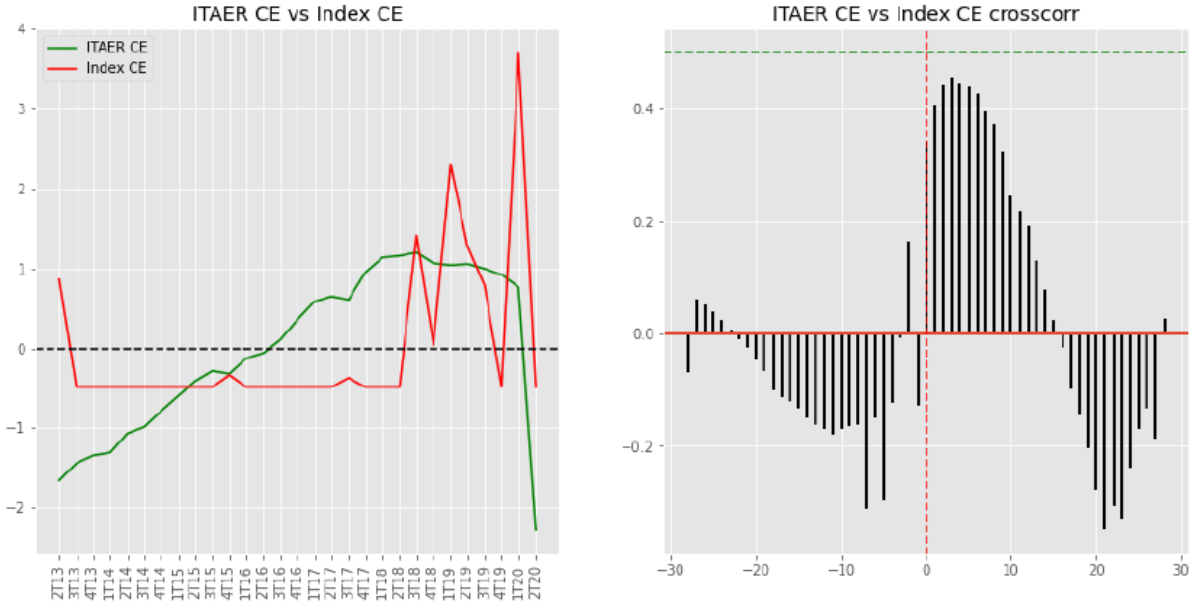
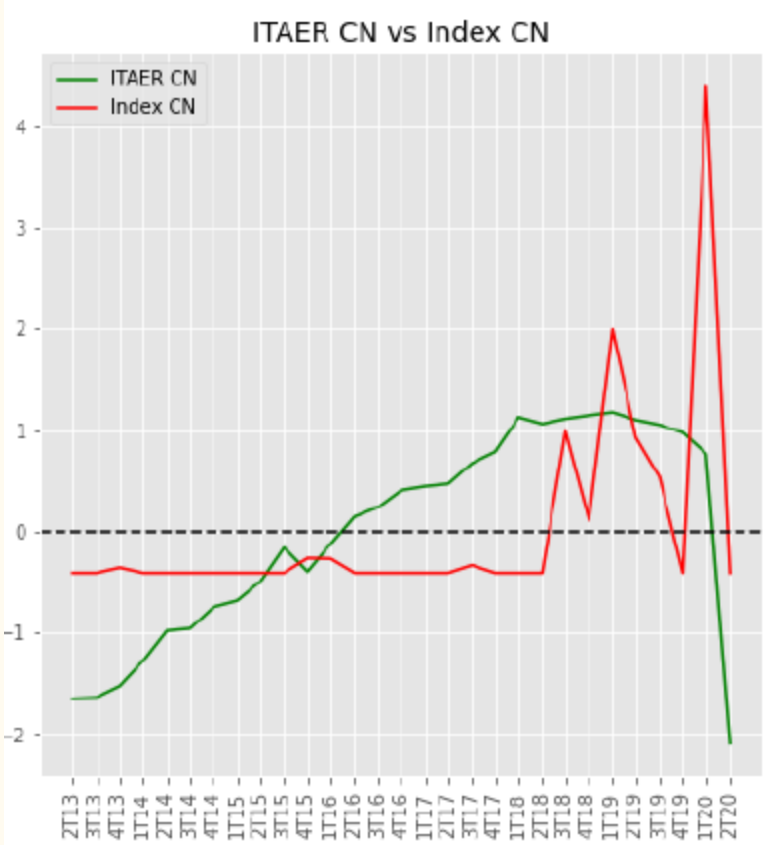
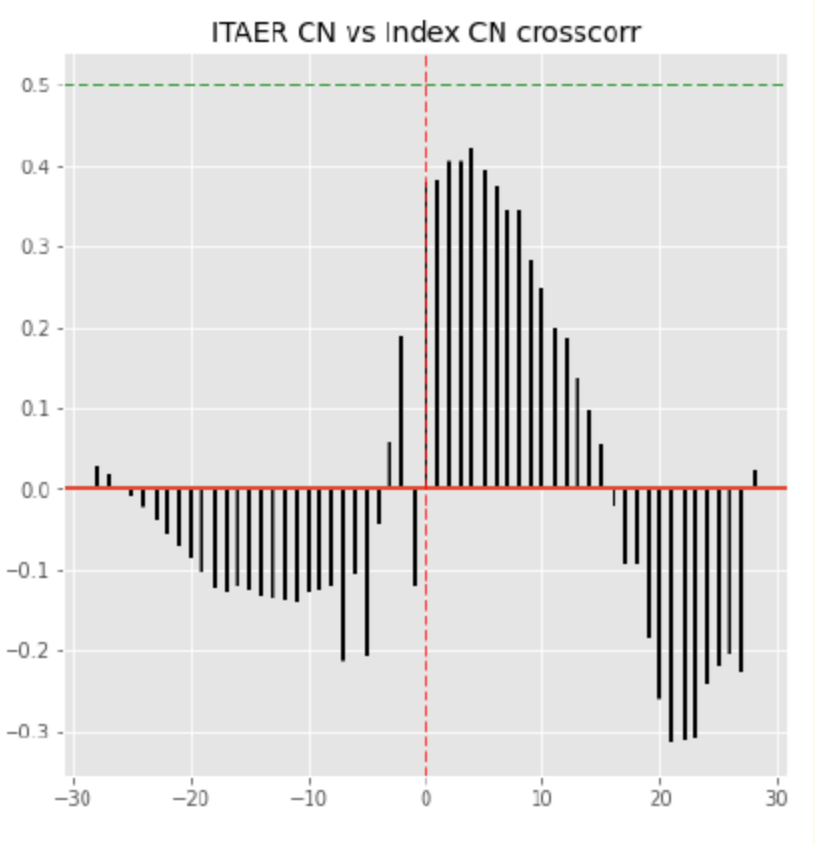


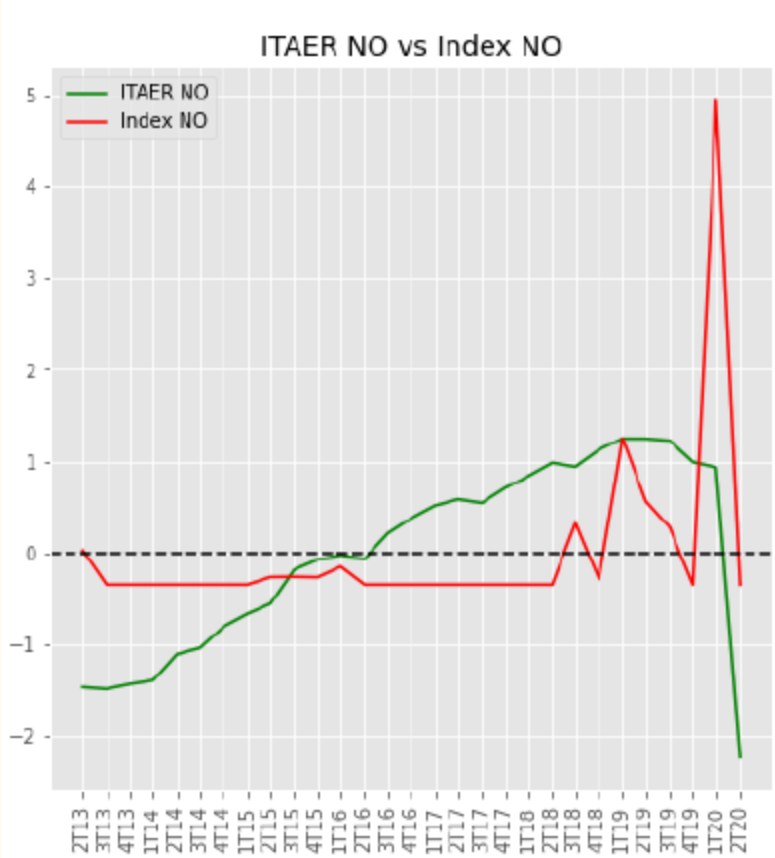
Figura 3.6: a) Indicadores regionales contra el **ITAER** del 2T13 al 2T20, b) correlación cruzada para el indicador y el **ITAER**.

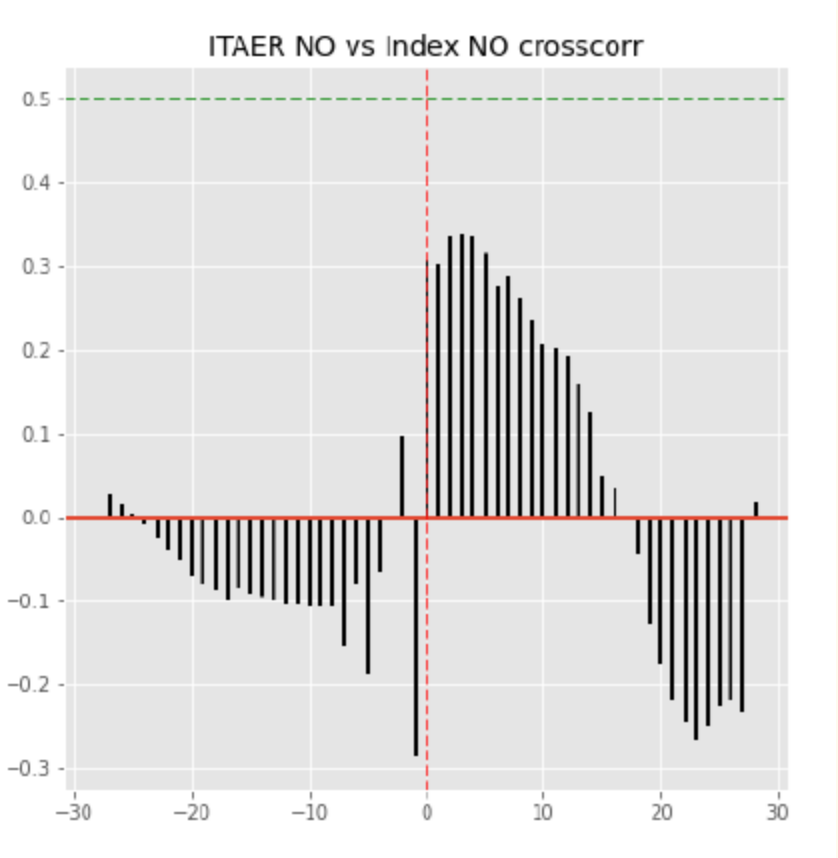
Indicador, Regional: CN



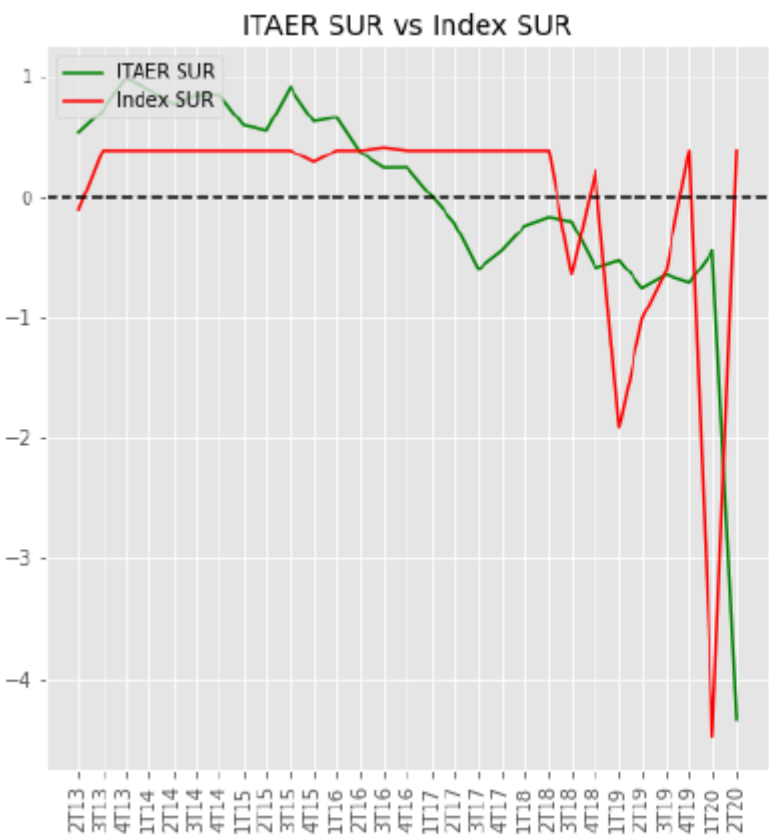


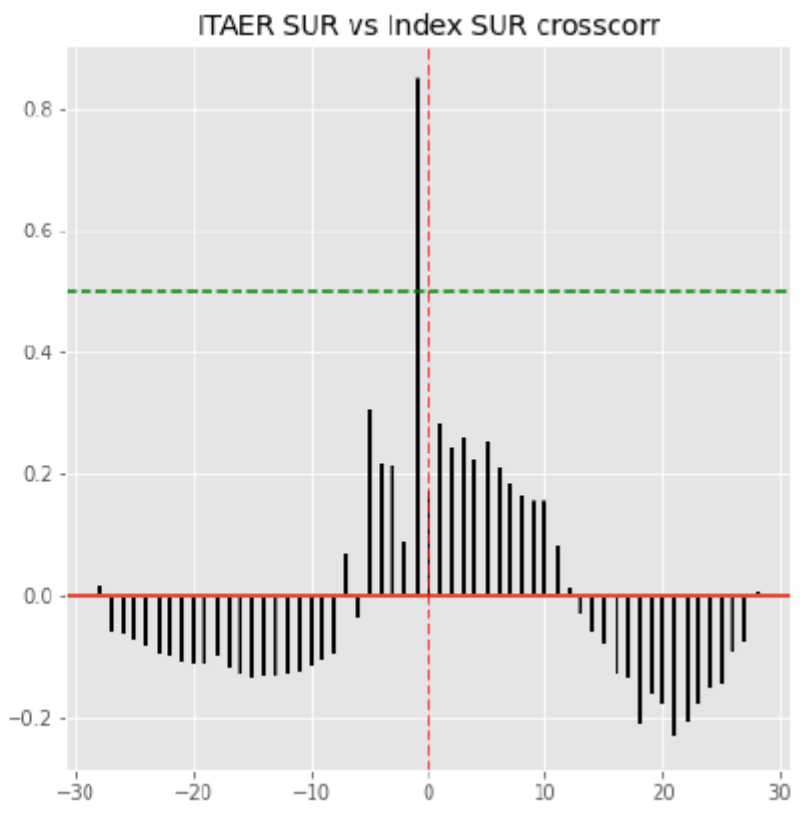
Indicador, Regional: NO



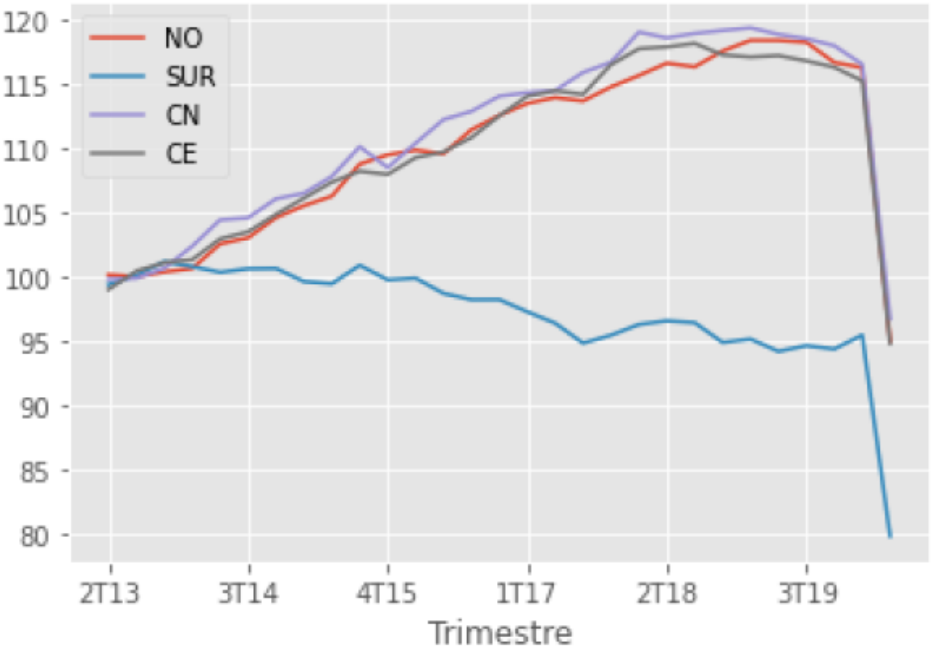


Indicador, Regional: SUR





Indicador, Regional



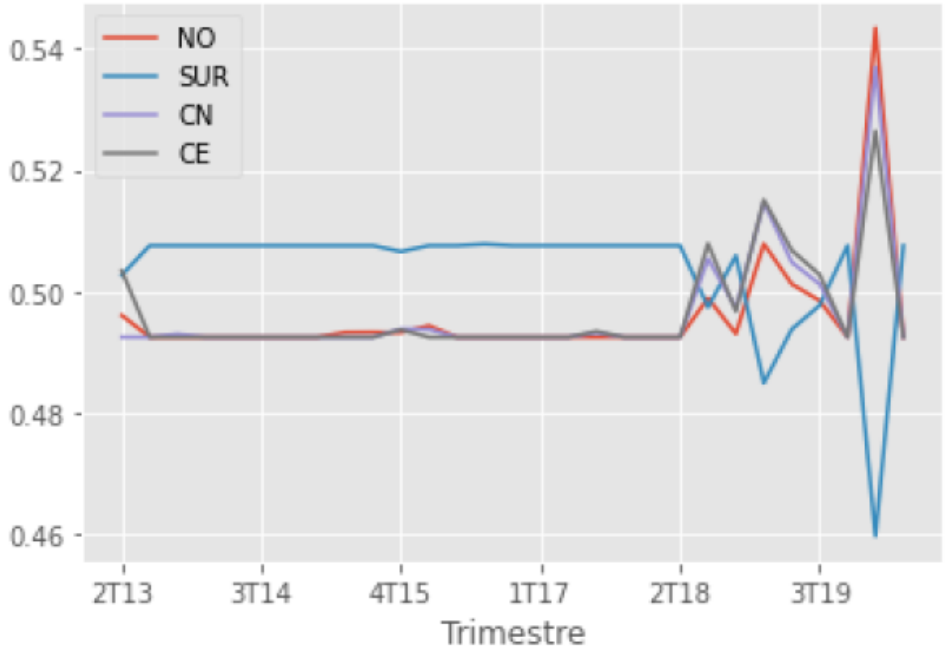


Figura 3.7: a) ITAER para cada región. b) Índice regional

Métodos basados en Machine Learning y Deep Learning para la clasificación de textos aplicado a las encuestas.

—

Problemática

El equipo de BANXICO se encuentra realizando una metodología para el correcto etiquetado de los textos provenientes de las encuestas del reporte sobre Economías Regionales, particularmente de los Riesgos a la alza y a la baja, así como de las demandas.

Una vez que se tengan los datos etiquetados, se desea poder encontrar algún método de Machine Learning o Deep Learning que permita obtener los criterios para una correcta clasificación de las encuestas futuras.

Propósito

Proponer diversas metodologías para la clasificación de textos, que puedan ser aplicadas posteriormente por el equipo de Banxico en las encuestas del reporte sobre Economías Regionales que se están etiquetando.

Particularmente, se proponen varias metodologías para la clasificación de las encuestas de acuerdo al sentimiento (Positivo, Negativo, Neutral).

Generación de etiquetas (sentimiento)

Textos de riesgos a la alza y a la baja 4T15 - 4T19

(Español)

Traducción de los textos*

(Inglés)

Asignación de sentimiento

Uso de Vader

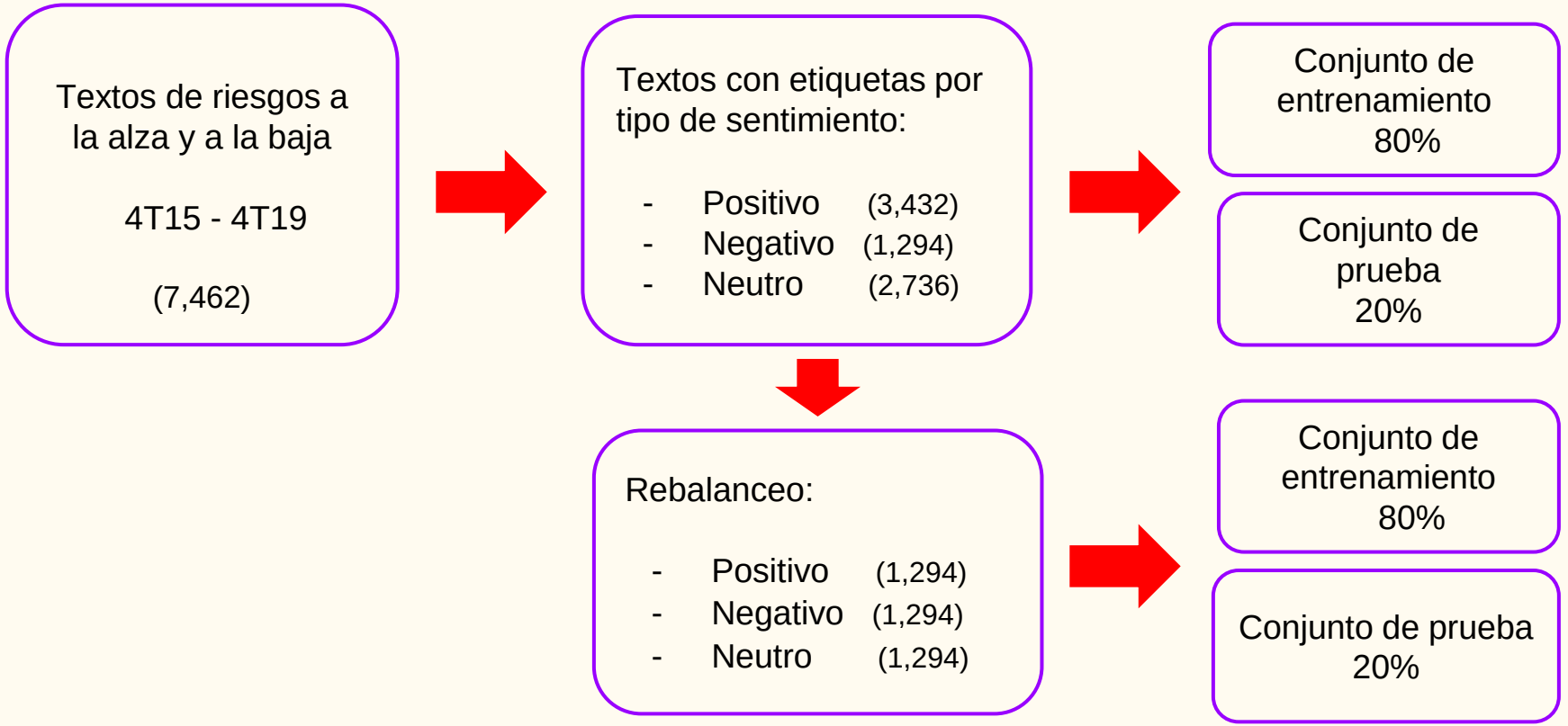
Métrica **Compound, C ∈ [0,1]:**

Positivo: C > 0.5
Negativo: C < -0.5
Neutro : e.o.c

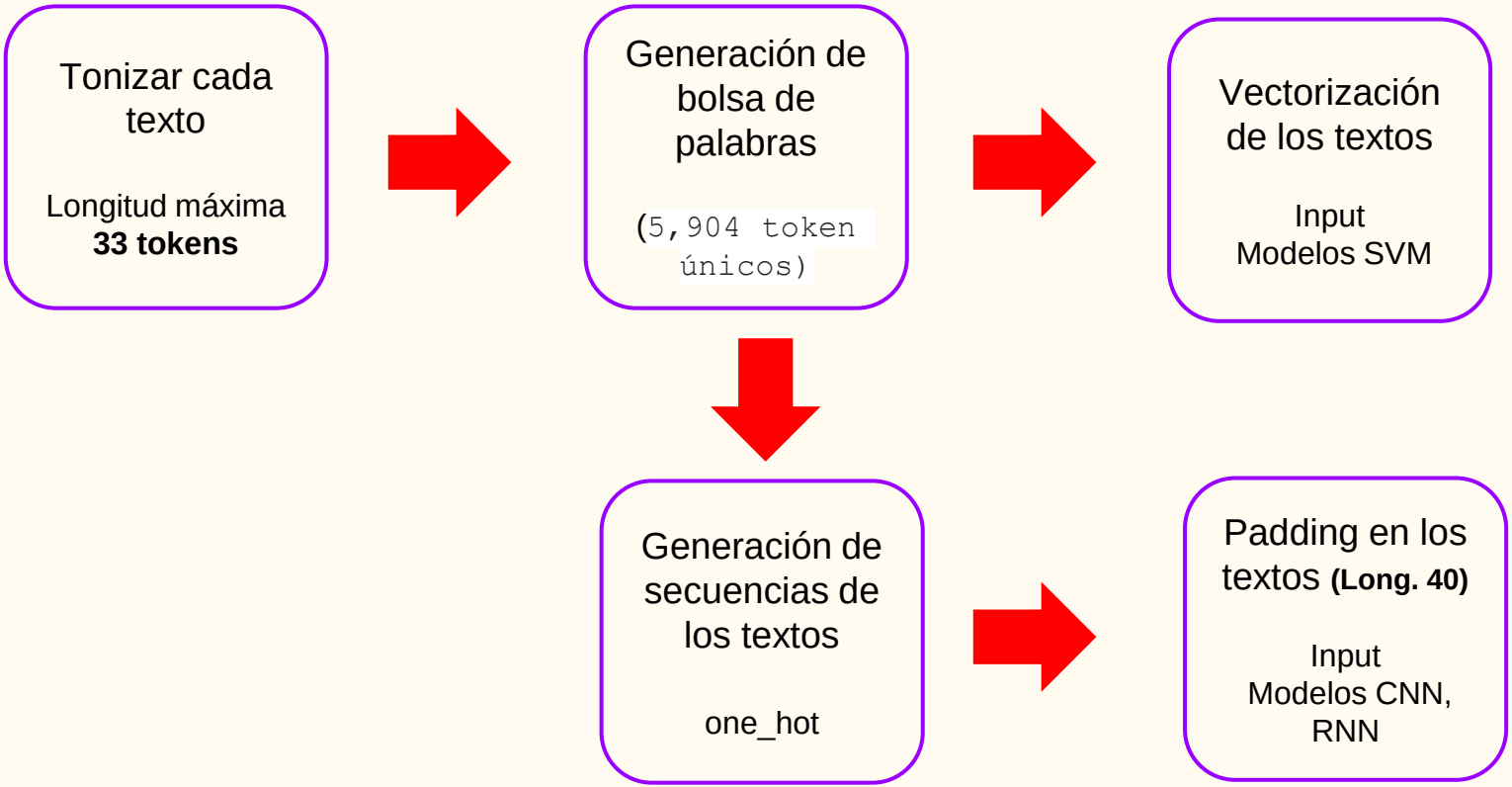
python + Google Translate

* Miranda, 2021

Generación de datos de entrenamiento y prueba



Tratamiento de los textos



Modelos de clasificación por sentimiento

Modelos propuestos

- Máquina de soporte vectorial (SVM)
- Redes neuronales convolucionales (CNN)
- Redes neuronales recurrentes (RNN)

1. Máquina de soporte vectorial (SVM)

Sin rebalanceo

SVM Modelo1 — Mode TD-IDF

Accuracy:			81.0%
Sentimiento	Precisión	Recall	F1-score
Neutral	0.77	0.85	0.81
Positivo	0.86	0.83	0.84
Negativo	0.76	0.67	0.71
Macro avg.	0.80	0.78	0.79

SVM Modelo2 — Mode count

Accuracy:			84.0%
Sentimiento	Precisión	Recall	F1-score
Neutral	0.79	0.92	0.85
Positivo	0.90	0.82	0.86
Negativo	0.83	0.71	0.76
Macro avg.	0.84	0.82	0.83

Con rebalanceo

SVM Modelo1 — Mode TD-IDF

Accuracy:			78.0%
Sentimiento	Precisión	Recall	F1-score
Neutral	0.70	0.85	0.77
Positivo	0.77	0.75	0.76
Negativo	0.88	0.73	0.80
Macro avg.	0.78	0.78	0.78

SVM Modelo2 — Mode count

Accuracy:			81.0%
Sentimiento	Precisión	Recall	F1-score
Neutral	0.73	0.92	0.81
Positivo	0.85	0.75	0.80
Negativo	0.88	0.77	0.82
Macro avg.	0.82	0.81	0.81

$$precisión = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

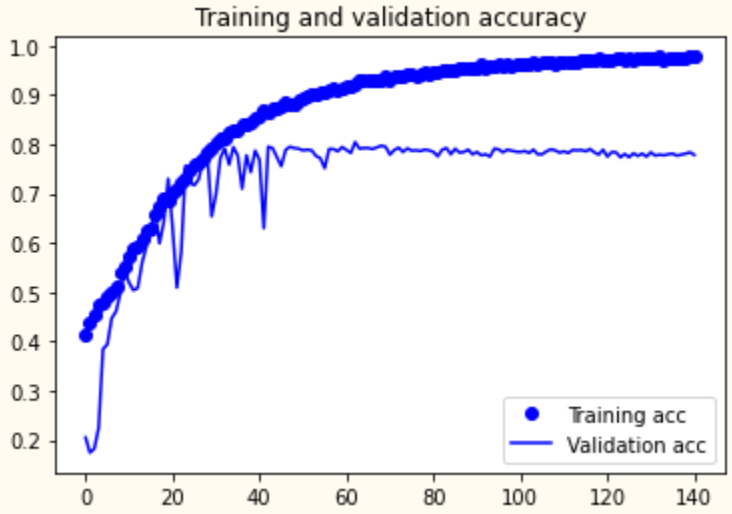
2. Redes neuronales convolucionales (CNN)

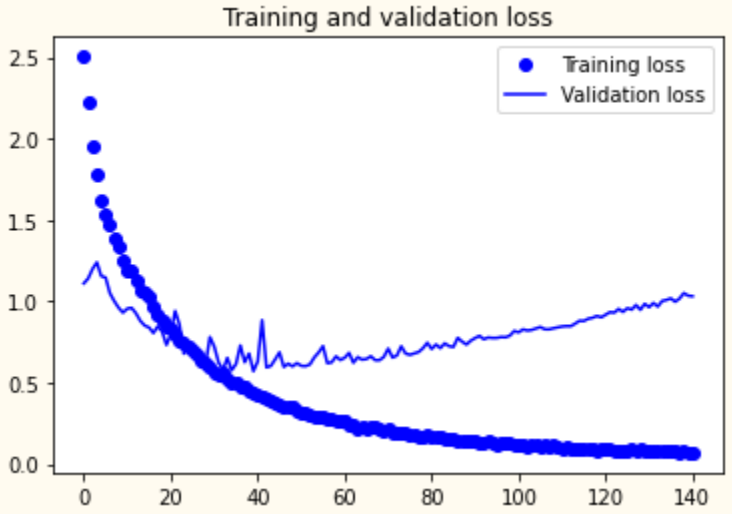
```
model = Sequential()
model.add(layers.Embedding(input_dim=vocab_size, output_dim=64))
model.add(layers.Conv1D(64, 7, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(layers.MaxPooling1D(2))
model.add(layers.Conv1D(32, 7, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(layers.GlobalMaxPooling1D())
model.add(layers.Dense(3, activation='softmax'))

EarlyStopping( monitor='acc', patience=8),
model.compile(optimizer=RMSprop(lr=1e-4),loss='categorical_crossentropy',metrics=['acc'])

model.fit(padded_docs_train, tr_y_new,epochs=500, batch_size=64,callbacks=callbacks_list,
                    validation_data=(padded_docs_test,te_y_new),verbose=1)
```

Redes neuronales convolucionales (CNN)

	Sin rebalanceo			Con rebalanceo		
CNN Modelo1 optimizador RMSprop	Accuracy:		77.8%	Accuracy:		76.3%
	Sentimiento	Precisión	Recall	F1-score	Precisión	Recall	F1-score
	Neutral	0.80	0.76	0.78	0.76	0.79	0.77
	Positivo	0.81	0.82	0.82	0.76	0.73	0.75
	Negativo	0.64	0.71	0.68	0.77	0.77	0.77
	Macro avg.	0.75	0.76	0.76	0.76	0.76	0.76
CNN Modelo2 optimizador adam	Accuracy:		77.5%	Accuracy:		74.0%
	Sentimiento	Precisión	Recall	F1-score	Precisión	Recall	F1-score
	Neutral	0.80	0.74	0.77	0.75	0.72	0.73
	Positivo	0.80	0.85	0.82	0.75	0.70	0.72
	Negativo	0.64	0.77	0.77	0.73	0.80	0.76
	Macro avg.	0.75	0.75	0.75	0.74	0.74	0.74

Redes neuronales convolucionales (CNN)





3. Redes neuronales recurrentes (RNN)

```
model = Sequential()
model.add(layers.Embedding(input_dim=vocab_size, output_dim=128))
model.add(LSTM(128, return_sequences=True))
model.add(Dropout(0.5))
model.add(LSTM(128, return_sequences=True))
model.add(Dropout(0.5))
model.add(LSTM(64))
model.add(Dropout(0.5))
model.add(Dense(64))
model.add(BatchNormalization())
model.add(Dense(32))
model.add(BatchNormalization())
model.add(Dense(32))
model.add(Dense(3, activation='softmax'))

EarlyStopping( monitor='acc', patience=8),
model.compile(optimizer=RMSprop(lr=1e-4),loss='categorical_crossentropy',metrics=['acc'])
model.fit(BOW_train, tr_y_new, epochs=500,batch_size=64, callbacks=callbacks_list,
validation_data=(BOW_test,te_y_new),verbose=1)
```

Redes neuronales recurrentes (RNN)

RNN Modelo1 — optimizador RMSprop

Sin rebalanceo			
Accuracy:	77.9%		
Sentimiento	Precisión	Recall	F1-score
Neutral	0.78	0.75	0.77
Positivo	0.78	0.85	0.82
Negativo	0.75	0.64	0.69
Macro avg.	0.77	0.75	0.76

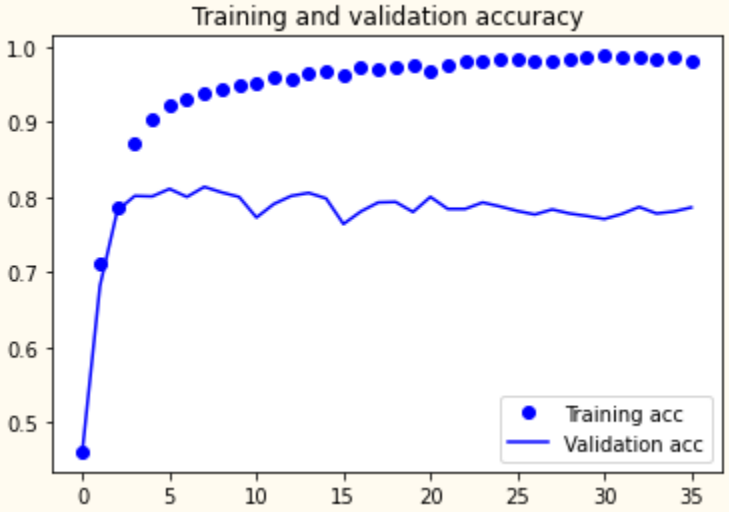
Con rebalanceo			
Accuracy:	72.3%		
Sentimiento	Precisión	Recall	F1-score
Neutral	0.66	0.73	0.69
Positivo	0.78	0.68	0.72
Negativo	0.74	0.76	0.75
Macro avg.	0.73	0.72	0.72

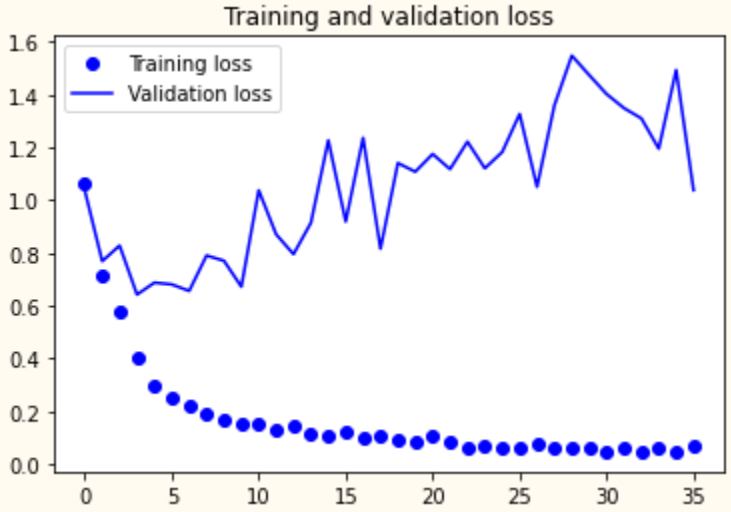
RNN Modelo2 — optimizador adam

Sin rebalanceo			
Accuracy:	78.6%		
Sentimiento	Precisión	Recall	F1-score
Neutral	0.80	0.78	0.79
Positivo	0.83	0.82	0.83
Negativo	0.66	0.70	0.68
Macro avg.	0.76	0.75	0.76

Con rebalanceo			
Accuracy:	75.3%		
Sentimiento	Precisión	Recall	F1-score
Neutral	0.68	0.82	0.74
Positivo	0.81	0.71	0.76
Negativo	0.79	0.74	0.76
Macro avg.	0.76	0.75	0.75

Redes neuronales recurrentes (RNN)





Conclusiones

- Se observó que **los cambios en el indicador de riesgo** están **fuertemente relacionado a los tópicos** de los que hablan, por ejemplo, cuando **hay cambios de gobierno, o con la pandemia actual** y es principalmente **influenciado por la componente del clúster 0** y muy poco por la del clúster 1.

- Además, se observó que **los indicadores de riesgo se relacionan a nivel general** con el **PIB de forma adelantada**, mientras que **de forma regional** con el **ITAER** de **forma adelantada** para la **región SUR** y para las **demás regiones de forma atrasada**.

Conclusiones

- Una razón para lo anterior es que **las otras regiones siguen la tendencia del clúster 2**, el cual contiene **tópicos** relacionado a empresa y gobierno, pero **en un contexto un poco más positivo**.

- **Otra forma de interpretar esto**, es que **para el cluster 2** que es donde aparentemente están clasificándose la mayor parte de los etiquetados como riesgos al alza, **en realidad representan aquellos eventos que más les preocupan**, y por lo tanto vemos esas variaciones en el indicador, **es decir que cuando aumenta la componente 2, cae el indicador**.

Conclusiones

Se recomienda que las preguntas se hagan de tal forma que se fuerce a los **encuestados que contesten cosas diferentes**, es decir **que los temas sean excluyentes basados en los temas que ya contestaron**.

Ejemplos correctos de la base:

- riesgo a la **baja**: “*Que el Covid 19 afecte de manera prolongada la actividad”*,
- riesgo a la **alza**: “*Mejoras en tema de seguridad”*;

Ejemplos incorrectos de la base:

- riesgo a la **baja**: “*Efecto prolongado e intenso de la pandemia en México”*,
- riesgo al **alza**: “*Efecto moderado y temporal del problema de coronavirus”*

Conclusiones

- De esta forma **se podría capturar mejor la perspectiva** que tienen **del riesgo** basado en los tópicos de lo que están hablando.

Conclusiones

- Con respecto a la clasificación de textos por sentimiento, pese a que se espera que haya mejores resultados con los datos balanceados, en este caso no se vio reflejado debido a la poca cantidad de instancias para el entrenamiento de los modelos.

- Sin embargo, aún con pocos datos, el modelo Máquina de soporte vectorial resulta ser el que mejor clasifica los textos de acuerdo al sentimiento, con un 84% de exactitud y mediante una vectorización de los textos por conteos.

Conclusiones

- Finalmente, es importante mencionar que los resultados de los modelos de clasificación mostrados, son resultado de una parametrización particular, la cual puede ser mejorada una vez que se tenga la tarea específica a resolver y las encuestas completas etiquetadas, para buscar las parametrizaciones adecuadas.

Trabajo futuro

- Replicar las 3 metodologías de clasificación de textos propuestas, generando los datos de entrenamiento y prueba a partir de los textos etiquetados por el equipo de BANXICO del periodo 2011 - 2020 y buscando una parametrización adecuada de acuerdo al criterio en que se hayan etiquetado los textos.

Gracias por su atención.