

MASTER UNIVERSITARIO



TRABAJO FIN DE MÁSTER

GENERACIÓN AUTOMÁTICA DE TIMELINES CON TRANSFORMERS

Alumno: **BEGOÑA LÓPEZ RODRÍGUEZ**

Director: **SERGIO LUIS NÁÑEZ ALONSO**

Convocatoria:

Año: 2021

o

AGRADECIMIENTOS

A mi familia, siempre, a mis compañeros de empresa, Ayesa, que me han apoyado y facilitado la realización de este Máster y a los profesores que he tenido y que tanto me han aportado e ilusionado a aprender y poder contribuir más a la sociedad. Agradecimiento especial a Sergio, mi director de TFM, por cómo nos hace estructurarlo y plantearlo desde el inicio.

En este trabajo fin de máster he abordado un reto complejo que me ha llevado a explorar varios caminos, muchos de ellos frustrantes en los resultados que iba obteniendo. El contar con una metodología de trabajo desde el inicio me ha permitido no perderme y continuar avanzando con una hoja de ruta clara y definida.

He aprendido muchísimo de NLP y, sobre todo, de cómo se abordan trabajos de investigación. Así que no puedo más que dar las gracias a todos los que me han ayudado a hacerlo posible

ÍNDICE

Agradecimientos	1
Índice	2
Índice de Figuras Y TABLAS.....	4
Abreviaturas	5
Introducción.....	6
Objetivos.....	8
Capítulo I. Material y metodología	10
1. Elección del dataset	10
1.1. Requisitos y condicionantes	10
1.2. Selección del dataset.....	11
2. Metodología de trabajo.....	12
2.1. Entendimiento del problema.....	12
2.2. Estudio del Estado del arte	12
2.3. Análisis de alternativas	12
Capítulo II. Creación del dataset.....	14
1. Obtención de los datos.....	14
1.1. Generación de la colección de urlS a partir de Twitter.....	14
2. Limpieza y validación	18
Capítulo III. Entendimiento del problema.....	21
1. Entendimiento del problema.....	21
2. Diseño de la solución.....	22
Capítulo IV. Estudio del estado del arte.....	24
1. Identificación de tópicos. Topic modelling	24
2. Extracción de información relevante y resumen. Summarization.....	27
Capítulo V. Análisis de alternativas topic modelling	32
1. Funcionamiento Top2Vec	32
2. Alternativa 1: Top2Vec con BERT Sentence Transformer.....	33
3. Alternativa 2: Top2Vec con doc2vec	41
Capítulo V. Análisis de alternativas summarization	43
4. Alternativa 1: Concatenación de noticias y resumen abstractivo	50
4.1. Estrategia y diseño del modelo	50
4.2. Implementación.....	50
4.3. Análisis de resultados	53
5. Alternativa 2: Resumen abstractivo por noticia, concatenación y resumen abstractivo	54

5.1. Estrategia y diseño del modelo	54
5.2. Implementación	54
5.3. Análisis de resultados	55
6. Alternativa 3: Resumen extractivo por noticia, concatenación y resumen abstractivo	56
6.1. Estrategia y diseño del modelo	56
6.2. Implementación	56
6.3. Análisis de resultados	57
7. Alternativa 4: Resumen extractivo por noticia, concatenación y resumen extractivo	59
7.1. Estrategia y diseño del modelo	59
7.2. Implementación	59
7.3. Análisis de resultados	59
Resultados y discusión / Desarrollo de la argumentación	62
Conclusiones finales.....	68
Bibliografía	70
Anexos.....	72
1. Resultados alternativa 1	72
2. Resultados alternativa 2	77
3. Resultados alternativa 3	83
4. Resultados alternativa 4	89

ÍNDICE DE FIGURAS Y TABLAS

Tabla 1: Tabla periódica de tareas NLP	6
Figura 2: Tf-idf o term frequency-inverse document frequency.....	25
Figura 3: Probabilidad de ver un documento y un término de manera conjunta	25
Figura 4: Representación del espacio semántico de embeddings de doc. (morado) y palabras (verde)	26
Figura 5: Arquitectura de los modelos Transformers	29
Tabla 1: Resultados modelado de tópicos Top2Vec con BERT Sentence Transformer	37
Tabla 2: Resultados modelado de tópicos con Top2Vec con doc2vec	41
Tabla 3: Resultados alternativa 1	53
Tabla 4: Resultados alternativa 2	55
Tabla 5: Resultados alternativa 3	58
Tabla 6: Resultados alternativa 4	60
Tabla 7: Resultados Top2Vec con BERT Sentence Transformer.....	64
Tabla 8: Resultados Top2Vec con doc2vec	64
Tabla 9: Resultados alternativa 1 - CONCATENACIÓN DE NOTICIAS Y RESUMEN ABSTRACTIVO	66
Tabla 10: Resultado alternativa 2 - RESUMEN ABSTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN ABSTRACTIVO	66
Tabla 11: Resultados alternativa 3 – RESUMEN EXTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN ABSTRACTIVO	66
Tabla 12: Resultados alternativa 4 - RESUMEN EXTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN EXTRACTIVO	66

ABREVIATURAS

- **TFM:** Trabajo Fin de Máster.
- **IA:** Inteligencia Artificial.
- **NLP:** Natural Language Processing
- **NER:** Named Entity Recognition

INTRODUCCIÓN

Internet supuso un nuevo paradigma sobre cómo compartimos y consumimos información. Ha sido una disrupción sin precedentes en nuestros hábitos, planteando retos complejos que permitan explotar las posibilidades que ofrece el acceso casi ilimitado a la extensísima y variada cantidad de información que se genera cada segundo. Grandes volúmenes de datos altamente desestructurados, gran parte de ellos textos. La Inteligencia Artificial, concretamente el ámbito NLP, está siendo un habilitador clave en este nuevo paradigma. Esta rama de la Inteligencia Artificial permite que las computadoras puedan entender, interpretar, analizar e interactuar con el lenguaje. Tareas altamente complejas además por la variedad de idiomas oficiales que se hablan en el mundo, más de 7.000, a los que debemos sumar dialectos, jergas profesionales, sociales, regionales.

El NLP no es una ciencia reciente. No obstante, los nuevos avances tecnológicos, capacidad de computación, nuevos algoritmos... están haciendo de catalizadores para el desarrollo de aplicaciones relacionadas con la gestión de la información. ¿Cuáles son las principales tareas que ya, hoy en día, nos ayudan a consumir eficientemente información? Referencio un artículo reciente que Rob van Zoes publicaba en la revista digital Medium [1]. En él estructuraba a modo de tabla periódica las principales tareas que abordan en el ámbito del NLP.

Tabla 1: Tabla periódica de tareas NLP.

1 Bit Bits to Character Encoding											75 App Interactive App Creation														
2 Typ Manual Typewriting	8 Man Manual Annotation											29 Pri Price Parser											63 Nex Next Token Prediction	69 Rel Relation Extraction	76 Ann Annotated Text Visualization
3 Str Loading a Structured Dataset	9 Act Annotation with Active Learning	14 Tok Tokenization	19 Ste Stemming	24 Ngr N-grams	30 Geo Geocoding	www.innerec.com										43 Trn Training Models	48 Spa Spam Detection	53 Key Keyword Extraction	58 Syn Wordnet Synsets	64 Rep Report Writing	70 Qan Question Answering	77 Wcl Workload			
4 Cor Generating a Corpus	10 Pro Training Data Provider	15 Voc Vocabulary Building	20 Lem Lemmatization	25 Phr Rulebased Phrasematcher	31 Tmp Temporal Parser	35 Sen Sentencizer	39 Ded Deduplication	44 Tst Evaluating Models	49 Sed Sentiment and Emotion Detection	54 Esu Extractive Summarization	59 Dst Distance Measures	65 Tra Machine Translation	71 Cha Chatbot Dialogue	78 Emb Word Embedding Visualization											
5 Api Loading from API	11 Cro Crowdsourcing Marketplace	16 Mor Morphological Tagger	21 Nrm Normalization	26 Chu Dependency Nounchunks	32 Nel Named Entity Linking	36 Par Paragraph Segmentation	40 Raw Raw Text Cleaning	45 Exp Explaining Models	50 Int Intent Classification	55 Top Topic Modeling	60 Sim Document Similarity	66 Abs Abstractive Summarization	72 Sem Semantic Search Indexing	79 Tim Events on Timeline											
6 Scr Text and File Scraping	12 Aug Textual Data Augmentation	17 Pos Part-of-Speech Tagger	22 Spl Spell Checker	27 Ner Named Entity Recognition	33 Crf Conditional Resolution	37 Grm Grammar Checker	41 Met Meta-Info Extractor	46 Dpl Deploying Models	51 Cls Text Classification	56 Tre Trend Detection	61 Dis Distributed Word Representations	67 Prp Paraphrasing	73 Kno Knowledge Base Population	80 Map Locations on Geomap											
7 Ext Text Extraction and OCR	13 Rul Rulebased Training Data	18 Dep Dependency Parser	23 Neg Negation Recognizer	28 Abr Abbreviation Finder	34 Anm Text Anonymizer	38 Rea Readability Scoring	42 Lng Language Identification	47 Mon Monitoring Models	52 Mlc Multi-Label Multi-Class Classification	57 Out Outlier Detection	62 Con Contextualized Word Representations	68 Lon Long Text Generation	74 Edi E-Discovery	81 Gra Knowledge Graph Visualization											
Source Data Loading	Training Data Generation	Word Parsing	Word Processing	Phrases and Entities	Entity Enriching	Sentences and Paragraphs	Documents	Model Development	Supervised Classification	Unsupervised Signaling	Similarity	Natural Language Generation	Systems	Information Visualization											

Periodic Table of Natural Language Processing Tasks is created with the Periodic Table Creator

Fuente: Artículo de Rob van Zoes [1]

Cito a modo de ejemplo las aplicaciones más comunes que utilizamos de manera habitual en las aplicaciones web y apps y sin las que no sería posible realizar la mayor parte de acciones más cotidianas:

- Motores de búsqueda
- Clasificación de documentos. Categorización y análisis de sentimiento
- Conversión voz-texto y texto-voz
- Traducción
- Resumen y síntesis de textos
- Extracción de información contextual. Preguntas/respuestas
- Generación de texto. Análisis predictivo de la siguiente palabra

Cuando me planteé sobre qué tema quería enfocar mi Trabajo Fin de Máster (TFM) tuve claro que quería profundizar en las capacidades que existen actualmente en el ámbito NLP, e intentar ver si podía contribuir con alguna aplicación específica que mejorara nuestra experiencia cotidiana de interacción con la información digital. Soy consumidora habitual de prensa digital y siempre he echado en falta una herramienta que nos diera una visión global de qué hechos relevantes han ocurrido con respecto a un tema. Me parece que sería una aplicación muy útil para incorporar en los buscadores de los diferentes periódicos digitales y/o, incluso, construir con esta funcionalidad una app específica que indexara las noticias de varios medios digitales y te proporcionara una visión más global, concreta e imparcial. Me parecería especialmente útil si se implementara para distintos países, ya que proporcionaría una capacidad de contexto inmediata sobre cualquier tema que se quiera empezar a analizar en ámbitos de actualidad que puede que conozcamos poco.

OBJETIVOS

Defino los siguientes objetivos a abordar en el marco de mi TFM:

- **Diseño y desarrollo de un modelo/herramienta** que permita **identificar hechos relevantes** ocurridos respecto a un tema específico analizando un *dataset* de noticias digitales. Un **generador de Timelines**.
- **Estudio del estado del arte** de las principales funcionalidades desarrolladas en el ámbito del **NLP** y análisis de su aplicabilidad a la experiencia de interacción con los contenidos digitales
- **Estudio del estado del arte de modelos NLP entrenados en español**. La carrera de la IA la están liderando sin duda EEUU y China. Las capacidades en IA serán clave en los próximos años y Europa/España no pueden quedarse atrás. En lo que respecta al idioma, el hecho de no tener modelos específicos entrenados en español nos dejará limitados a lo que desarrollen otros, sin capacidad por tanto de iniciativa. Además, tendríamos que asumir un riesgo evidente en lo que respecta a la falta de control sobre aspectos tan críticos como son los relativos a la ética y el control de los sesgos

Capítulo I. Material y metodología

CAPÍTULO I. MATERIAL Y METODOLOGÍA

1. ELECCIÓN DEL DATASET

1.1. REQUISITOS Y CONDICIONANTES

El primer paso para poder comenzar el proyecto de **Generación Automática de Timelines** es disponer de un *dataset* adecuado. Imprescindible tanto para estudiar las diferentes funcionalidades de IA que decidamos explorar, como para, llegado el caso, poder desarrollar nuestro propio modelo. Este *dataset* debe reunir los siguientes **requisitos**:

- **Noticias digitales** con al menos los siguientes campos:
 - **Título**
 - **Headline**
 - **Contenido noticia**
 - **Fecha**
- Debe cubrir un **espacio temporal amplio** que permita que hayan ocurrido diversos hechos relevantes respecto a temas específicos de actualidad
- Debe cubrir **temáticas variadas**
- Debe ser de un medio que publique **nativamente en español**. Descartamos los *dataset* de medios digitales que publiquen en otros idiomas y que hayan sido traducidos automáticamente

Por otro lado, existen algunos **condicionantes** que debemos tener en cuenta:

- Seleccionaré preferentemente un *dataset* de **noticias publicadas en España en un horizonte temporal cercano**. El conocimiento de detalle de la actualidad reciente del lugar donde resido me ayudará sin duda a captar mejor los matices sobre cómo están funcionando los diferentes modelos
- Gran parte de los medios digitales tienen ya planes de suscripción para el acceso a sus contenidos. Me limitaré a los **medios con contenidos abiertos**

1.2. SELECCIÓN DEL DATASET

Para buscar qué *datasets* podría utilizar he accedido a los repositorios de Kaggle [3] y Hugging Face [4]. Tras una exhaustiva búsqueda he podido comprobar como el inglés es el gran idioma predominante en el ámbito NLP. Tras descartar algunos repositorios que se correspondían con noticias de medios en inglés traducidas de manera automática y un *dataset* pequeño de un medio español especializado en deportes (no quiero limitar mi proyecto a un ámbito tan específico), decido construir mi propio *dataset* desarrollando un módulo de *Web Scraping*. Tras analizar varias opciones considero que una buena alternativa a valorar puede ser:

Medio digital: “20 Minutos”

Periodo de tiempo: 01/01/2020 – 31/05/2021

2. METODOLOGÍA DE TRABAJO

Con objeto de trabajar de una manera estructurada y asegurar que doy los pasos adecuados para entender a qué retos debo dar respuesta y qué opciones tengo que analizar, planteo la siguiente metodología de trabajo

2.1. ENTENDIMIENTO DEL PROBLEMA

En una primera fase trabajaré en un análisis de concepto de qué problema o problemas tenemos que resolver. Es importante definir en detalle qué resultados queremos obtener (¿qué significa que un hecho sea relevante?), y desglosar o subdividir este reto en problemas individuales que consigan dichos resultados.

2.2. ESTUDIO DEL ESTADO DEL ARTE

Una vez identificados los diferentes problemas a resolver abordaremos un estudio del estado del arte de las diferentes tareas NLP implicadas en esas soluciones. Entre ellas estarán sin duda las de Resumen o *Summarization*, Modelado de Tópicos y puede que *Question/Answer*.

Con las técnicas de NLP recientes que considere más adecuadas para resolver nuestro reto, plantearé varias alternativas de estudio y/o modelado con objeto de poder comprobar su funcionamiento y obtener una comparativa.

2.3. ANÁLISIS DE ALTERNATIVAS

Abordaremos los siguientes pasos para cada alternativa que planteemos:

1.1.1. Estudio de la arquitectura y funcionamiento de la alternativa

Estudio teórico de la técnica NLP. Principios conceptuales, arquitectura y principales usos

1.1.2. Definición de estrategia y diseño del modelo

Planteamiento de la estrategia de solución y diseño del modelo

1.1.3. Implementación

Implementación, entrenamiento y pruebas

1.1.4. Análisis de resultados y evaluación

Análisis de los resultados y primera valoración de la alternativa

Capítulo II. Creación del dataset

CAPÍTULO II. CREACIÓN DEL DATASET

1. OBTENCIÓN DE LOS DATOS

Como ya se ha explicado en el apartado 1.2 del capítulo I del presente documento, decido abordar una tarea de *Web Scraping* para obtener un *dataset* de noticias en español lo suficientemente amplio en el tiempo que permita la construcción de un modelo de generación de timelines robusto. El periódico digital elegido es “20 minutos” y el periodo de tiempo 01/01/2020 – 31/05/2021.

Para la tarea de *Web Scraping* utilizo la librería *urllib*, las funciones *Request()* y *urlopen()* de esta librería me permiten acceder al contenido de una url que pasamos como parámetro. Necesito por tanto una colección de urls del periódico “20 minutos” que cubra el periodo de tiempo elegido con suficiente completitud. Analizo como posible fuente Twitter. En una primera aproximación de análisis selecciono varios días aleatorios y compruebo que estas publicaciones:

- Publican en torno a 200 tweets al día
- Casi el 90% de las publicaciones referencia la url de la noticia del periódico digital
- Las temáticas de actualidad de las publicaciones abarcan una amplia variedad de tópicos

Me decido por tanto a utilizar Twitter como fuente de direcciones url para montar mi *dataset*. Una vez construido y depurado, realizaré comprobaciones de completitud de los datos finales para asegurar que el *dataset* es válido.

1.1. GENERACIÓN DE LA COLECCIÓN DE URLS A PARTIR DE TWITTER

Para esta tarea utilizo en primera instancia la API de Twitter, haciendo uso de mi cuenta para desarrollos de integraciones con Twitter. El acceso a la API se realiza a través de la librería *tweepy*. Accedo a la documentación de la librería, [5] y [6], y configuro los parámetros principales. Realizo una prueba inicial de conexión con la API con mis credenciales y extraigo 200 tweets.

```
# Configuración de la API:
def twitter_config():
    # Autenticación y acceso usando claves:
    auth = tweepy.OAuthHandler("XXXX", "XXXX")
    auth.set_access_token("XXXX", "XXXX")

    # Devolvemos la utenticación:
    api = tweepy.API(auth)
    return api
```


Generación automática de timelines con Transformers

```
twt_api = twitter_config()

tweets = twt_api.user_timeline(screen_name="@20m", count=200)
print("Número de tweets extraídos: {}".format(len(tweets)))

Número de tweets extraídos: 200.

print("Ejemplo de 5 tweets:\n")
for tweet in tweets[(200-5):]:
    print(tweet.text)
    print()

Ejemplo de 5 tweets:

Rescatado por mar un menor de edad en coma etílico tras acceder a una cala con acceso prohibido en La Coruña https://t.co/4eDalk7c8Y

Kristen Stewart y Léa Seydoux se apuntan a los 'Crimes of the Future' de David Cronenberg https://t.co/SfLYu6A7MM

Eiza González encarnará a la legendaria actriz mexicana María Félix https://t.co/gz1fAiafjW

Un potencial secuestro en el golfo de Omán implica a varias embarcaciones https://t.co/ijFVaKKicr

Laura Matamoros desvela a quién se parece su bebé con un divertido mensaje de su madre: "Va veremos" https://t.co/iZcCns83K3
```

El tiempo de ejecución para la obtención de estos 200 tweets es mínimo, casi inmediato. Intento realizar una extracción más voluminosa con objeto de evaluar tiempos y me encuentro con el primer problema; la API de Twitter, tal y como se detalla en el apartado API Reference de su documentación [5], sólo te permite extraer los tweets de los últimos 30 días. Este problema descarta esta opción, *tweepy*, y analizo qué otras opciones existen para extraer tweets.

La limitación temporal de *tweepy* es un problema bien conocido y documentado en los foros de Python, en los que se recomienda utilizar, para extracciones más masivas, la librería *twint* [7].

```
import twint
#configuration
config = twint.Config()
config.Username = "20m"
config.Pandas = True
config.Since = "2020-01-01"
config.Until = "2021-05-31"
config.Output = "News.csv"

#running search
twint.run.Search(config)

C:\ProgramData\Anaconda3\lib\json\decoder.py:353: RuntimeWarning: coroutine 'Twint.main' was never awaited
obj, end = self.scan_once(s, idx)
RuntimeWarning: Enable tracemalloc to get the object allocation traceback

1399150721459757057 2021-05-31 01:48:21 +0200 <20m> Secuestran a unos 200 estudiantes de una escuela en Nigeria https://t.co/CIxS017K4y
1399148623347265539 2021-05-31 01:40:01 +0200 <20m> ¿Cuántas veces se debe usar una prenda de ropa entre lavado y lavado? https://t.co/rcwh0upRRu
1399146734354374665 2021-05-31 01:32:31 +0200 <20m> Montero, Belarra y Díaz piden que el derecho al aborto se ejerza de forma "segura, pública y gratuita" en todas las comunidades https://t.co/aEpWYj5X9m
1399145146046992387 2021-05-31 01:26:12 +0200 <20m> Suspenden la búsqueda sin hallar al segundo desaparecido en el Puerto de Castellón https://t.co/m42UFB8PL7
1399143592195760129 2021-05-31 01:20:01 +0200 <20m> El truco de Karlos Arguiñano para hacer el huevo frito perfecto https://t.co/zb3UpwoCNS
1399141689701380100 2021-05-31 01:12:28 +0200 <20m> El derrumbamiento de una granja en Ourense deja decenas de cerdos atrapados bajo los escombros https://t.co/aoeacgFCOH
1399140105810550784 2021-05-31 01:06:10 +0200 <20m> Las presuntas estafas con criptomonedas inundan la Audiencia Nacional https://t.co/BxqHqIfOis
1399138550463002224 2021-05-31 01:00:01 +0200 <20m> El pasado oculto en las etiquetas de Zara que ha sido desvelado por sus
```

Estas instrucciones me permiten obtener un conjunto de **124.582 tweets** comprendidos entre el 01/01/2020 y el 31/05/2021, lo que nos da una media de **244,3 tweets diarios**. Una vez hayamos completado y depurado los datos, realizaremos un estudio de completitud temporal de los mismos, pero inicialmente, parece que estamos avanzando por el camino correcto.

Con objeto de manejar los datos en un modo más eficiente, los paso a un *dataframe* de la librería Pandas.

```
news_data=twint.output.panda.Tweets_df[["id", "date", "username", "tweet", "nlikes"]]
```

```
news_data[:10]
```

	id	date	username	tweet	nlikes
0	1399150721459757057	2021-05-31 01:48:21	20m	Secuestran a unos 200 estudiantes de una escue...	9
1	1399148623347265539	2021-05-31 01:40:01	20m	¿Cuántas veces se debe usar una prenda de ropa...	5
2	1399146734354374665	2021-05-31 01:32:31	20m	Montero, Belarra y Díaz piden que el derecho a...	9
3	1399145146046992387	2021-05-31 01:26:12	20m	Suspenden la búsqueda sin hallar al segundo de...	0
4	1399143592195760129	2021-05-31 01:20:01	20m	El truco de Karlos Arguiñano para hacer el hue...	12
5	1399141689701380100	2021-05-31 01:12:28	20m	El derrumbamiento de una granja en Ourense dej...	9
6	1399140105810550784	2021-05-31 01:06:10	20m	Las presuntas estafas con criptomonedas inunda...	13
7	1399138559462883334	2021-05-31 01:00:01	20m	El mensaje oculto en las etiquetas de Zara que...	15
8	1399136769904451586	2021-05-31 00:52:55	20m	Al menos dos muertos y 25 heridos por un tirot...	3
9	1399135169118883844	2021-05-31 00:46:33	20m	Detienen a un anciano sospechoso de haber atra...	8

A continuación, extraigo la url que referencia a la noticia del periódico digital.

```
#Extraemos la url de la noticia del tweet y la almacenamos en un campo nuevo
news_data['url']=news_data['tweet'].apply(lambda x: x[x.find('https'):])
news_data['url']=news_data['url'].apply(lambda x: x.replace(x[x.find(' '):], '') if x.find(' ') != -1 else x[x.find('https'):])
```

```
news_data[:10]
```

	id	date	username	tweet	nlikes	url
0	1399150721459757057	2021-05-31 01:48:21	20m	Secuestran a unos 200 estudiantes de una escue...	9	https://t.co/C1xS0i7K4y
1	1399148623347265539	2021-05-31 01:40:01	20m	¿Cuántas veces se debe usar una prenda de ropa...	5	https://t.co/rcwh0upRRu
2	1399146734354374665	2021-05-31 01:32:31	20m	Montero, Belarra y Díaz piden que el derecho a...	9	https://t.co/aEpWYj5X9m
3	1399145146046992387	2021-05-31 01:26:12	20m	Suspenden la búsqueda sin hallar al segundo de...	0	https://t.co/m42UF8PL7
4	1399143592195760129	2021-05-31 01:20:01	20m	El truco de Karlos Arguiñano para hacer el hue...	12	https://t.co/zbJUUpwoCNs
5	1399141689701380100	2021-05-31 01:12:28	20m	El derrumbamiento de una granja en Ourense dej...	9	https://t.co/aoeacgFCOH
6	1399140105810550784	2021-05-31 01:06:10	20m	Las presuntas estafas con criptomonedas inunda...	13	https://t.co/BxqHqIfOis
7	1399138559462883334	2021-05-31 01:00:01	20m	El mensaje oculto en las etiquetas de Zara que...	15	https://t.co/cSztidK6YX7
8	1399136769904451586	2021-05-31 00:52:55	20m	Al menos dos muertos y 25 heridos por un tirot...	3	https://t.co/k5yq3P1af6
9	1399135169118883844	2021-05-31 00:46:33	20m	Detienen a un anciano sospechoso de haber atra...	8	https://t.co/GHV5plutNX

En este punto, tendría todo lo necesario para abordar la tarea de *Web Scraping*. En primer lugar, decido qué campos voy a querer extraer de las diferentes noticias digitales, y estudio en qué parte del código HTML de dichas páginas se encuentra codificada dicha información. Los campos que decido extraer son:

- Título
- Contenido de la noticia
- Fecha

Creo campos nuevos en mi *dataframe*, inicialmente vacíos, para ir almacenando esta información conforme la vaya recuperando.

```
import numpy as np
news_data['titulo']=np.nan
news_data['headline']=np.nan
news_data['contenido']=np.nan
news_data['fecha']=np.nan
news_data['tags']=np.nan
```

Construimos finalmente el módulo de *Web Scraping*. Utilizamos las siguientes librerías:

- `urllib.request`. Accede al contenido de una url que se pasa como parámetro
- `http.cookiejar`. Nos permite gestionar las cookies
- `BeautifulSoup`. Lee el formato html
- `time.sleep`. Lo utilizamos para introducir retardos entre petición y petición y no sobrecargar el sitio web

```
import time
import random
from http.cookiejar import CookieJar
from urllib.request import urlopen, Request
import requests
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2228.0 Safari/537.3'}
from bs4 import BeautifulSoup
for i in range(0, len(news_data)):
    try:
        req = Request(news_data['url'][i], headers=headers)
        f_html=urlopen(req)
        long_res = BeautifulSoup(f_html.read(), "html5lib")
        long_url = long_res.select_one("title").get_text()
        html = urlopen(long_url)
        time.sleep(3*random.random())
        res = BeautifulSoup(html.read(), "html5lib")
        news_data['titulo'][i]=res.select_one("section div h1").get_text()
        content= res.find_all("article", class_="article-body")
        x = content[0].find_all('p')
        # Unificamos los párrafos
        list_paragraphs = []
        for p in np.arange(0, len(x)):
            paragraph = x[p].get_text()
            list_paragraphs.append(paragraph)
            final_article = " ".join(list_paragraphs)
            news_data['contenido'][i]=final_article
            date=res.find_all("span", class_="article-date")
            news_data['fecha'][i]=date[0].get_text()
    except:
        print('error:', news_data['url'][i])
```

Obtengo errores en aproximadamente el 9% de los registros. Tras una amplia revisión de las causas que provocan estos errores, el 100% se deben a referencias esporádicas a urls de otros periódicos digitales que tiene por tanto un formato HTML diferente al de “20 minutos”, que es el que he configurado. No son noticias significativas para este trabajo, por lo que decido obviarlos.

Una vez eliminados los errores, obtenemos un **primer dataset con 113.207 registros completo**.

```
#Eliminamos las columnas que se han añadido en el proceso de guardar/leer csv porque no nos aportan
news_completed=news_data.loc[:, ['id', 'date', 'username', 'nlikes', 'url', 'titulo', 'contenido', 'fecha']]
news_completed.head()
```

	id	date	username	nlikes	url	titulo	contenido	fecha
0	1399150721459757057	2021-05-31 01:48:21	20m	9	https://t.co/CixS0I7K4y	Secuestran a unos 200 estudiantes de una escue...	Unos 200 estudiantes de una escuela islámica d...	31.05.2021 - 01:37h
1	1399148623347265539	2021-05-31 01:40:01	20m	5	https://t.co/rcwh0upRRu	¿Cuántas veces se debe usar una prenda de ropa...	Poner la lavadora es una de las tareas domésti...	30.05.2021 - 19:31h
2	1399146734354374665	2021-05-31 01:32:31	20m	9	https://t.co/aEpWYj5X9m	Montero, Belarra y Díaz piden que el derecho a...	Las ministras de Igualdad, Irene Montero, la d...	30.05.2021 - 23:35h
3	1399145146046992387	2021-05-31 01:26:12	20m	0	https://t.co/m42UFB8PL7	Suspenden la búsqueda sin hallar al segundo de...	El operativo de búsqueda del estibador desapar...	30.05.2021 - 21:45h
4	1399143592195760129	2021-05-31 01:20:01	20m	12	https://t.co/zbJUUpwoCNs	El truco de Karlos Arguiñano para hacer el hue...	Saber cocinar un huevo frito es todo un básico...	31.05.2021 - 12:46h

```
len(news_completed)
```

```
113207
```

2. LIMPIEZA Y VALIDACIÓN

Una vez hemos obtenido un primer *dataset* completo, hacemos un primer análisis visual para comprobar si es necesario realizar alguna tarea de limpieza. Tras una inspección visual del detalle de los datos no apreciamos errores de formato, pero observamos que se repiten con bastante frecuencia noticias. Comprobamos que la repetición de los textos es idéntica, tanto en el título como en contenido. Suelen ser noticias que comienzan con “DIRECTO: ...”, debido a un formato específico de noticias en las que se va haciendo una publicación continuada como actualización de un tema de actualidad o evento concreto que se está siguiendo. Estas repeticiones desvirtuarían cualquier modelo que diseñe para la identificación de tópicos por lo que decido eliminar cualquier registro que esté duplicado.

```
#Eliminamos los registros duplicados
news_data.drop_duplicates('titulo', keep='last', inplace=True)
```

```
len(news_data)
```

```
95008
```

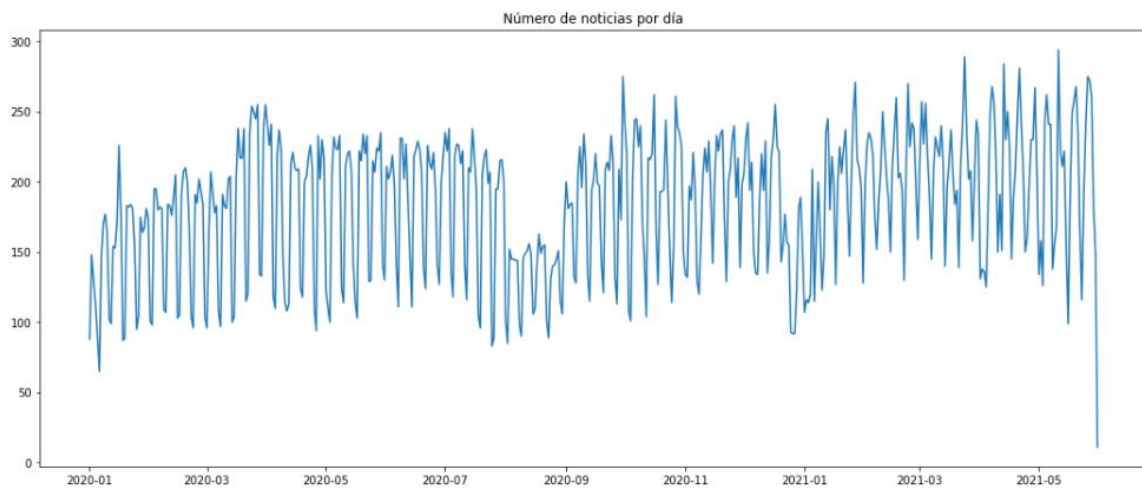
Nuestro nuevo *dataset* tiene 95.008 registros. Realizamos ahora un análisis para verificar la completitud de los datos.

Agrupamos por días y calculamos el total de noticias para ese día. Ordenado de mayor a menor:

```
#Agrupo por días para ver la completitud temporal de las noticias
a=pd.to_datetime(news['date']).dt.floor('d').value_counts()
a
```

```
2021-05-11    294
2021-03-24    289
2021-04-13    284
2021-04-21    281
2020-09-30    275
...
2020-01-18     87
2020-08-02     85
2020-07-25     83
2020-01-06     65
2021-05-31     11
Name: date, Length: 517, dtype: int64
```

Comprobamos todos los días tenemos noticias, en un volumen además importante. Salvo el 31-05-2021 (día que realizamos la extracción y, por tanto, no está completo), estamos siempre muy por encima de las 65 noticias diarias. Visualizamos gráficamente esta distribución para tener una imagen más completa.



Damos por válido el *dataset* y comenzamos nuestro trabajo de diseño y modelado.

Capítulo III.

Entendimiento del problema

CAPÍTULO III. ENTENDIMIENTO DEL PROBLEMA

1. ENTENDIMIENTO DEL PROBLEMA

El reto que nos hemos planteado abordar en este TFM consiste en desarrollar un modelo que analice un *dataset* de noticias e identifique para una temática concreta qué hechos relevantes han ocurrido. Partiendo de la premisa de que considerar que un hecho sea relevante es conceptualmente subjetivo, vamos a intentar objetivar, en base a la información de la que disponemos, cómo podemos catalogar la relevancia.

Nuestra única fuente de información es un *dataset* con noticias publicadas por un medio digital. ¿Cómo impacta en la publicación de noticias un acontecimiento relevante? Podemos realizar las siguientes afirmaciones:

- Se producirá un **pico de publicaciones** que hablen de ese hecho. Es decir, un incremento de las noticias que hablan de esa temática en un slot temporal específico
- Encontraremos **noticias** que estarán **más enfocadas** a explicar esa temática **y otras que**, partiendo de ella como base, **añadirán otras informaciones** relativas a la temática que pueden ser **más o menos difusas** y variadas

Estructurando este enfoque del problema en tareas concretas llegamos al siguiente desglose:

- Debemos **clasificar las diferentes publicaciones en “temáticas”**. Es decir, tenemos que ser capaces de entender qué hilos o conversaciones diferentes se están explicando, partiendo del hecho de que una misma publicación tratará con total seguridad sobre diferentes *hilos* o tópicos.
- Una vez identificados los tópicos, tenemos que **catalogarlos en relevantes/no relevantes, determinando además qué conjunto de noticias** describen mejor el acontecimiento que ha provocado el aumento de publicaciones
- Por último, tenemos que ser capaces de **resumir** este hecho a partir de dichas publicaciones

2. DISEÑO DE LA SOLUCIÓN

Las tareas del ámbito NLP que nos ayudarán a resolver las diferentes cuestiones identificadas son:

- **Topic modelling o modelado de tópicos.** De resolver el problema de identificación de temáticas, y clasificación de documentos que tratan dichas temáticas, se encarga el modelado de tópicos. Son modelos que tratan de identificar los diferentes tópicos o hilos que hay en un conjunto de documentos (pueden ser artículos científicos, noticias, chats, conversaciones...), y clasifica los documentos asignando un peso según la predominancia que tenga cada tópico en dicho documento. Con una relación *many-to-many*. Un tópico estará presente en más de un documento y un documento tratará sobre más de un tópico.
- **Summarization o resumen.** Una vez identificados los tópicos, y las noticias que centran o mejor describen el hecho relevante, tendremos sin duda que resumirlo, extrayendo la información clave que se comenta en dichas publicaciones

Diseñamos la siguiente solución:

- 1) Recuperación de documentos o *text retrieval*. **El primer paso será seleccionar las publicaciones relativas a la temática que se indique por parte del usuario.** El ámbito del *text retrieval*, como podemos comprobar fijándonos en la sofisticación de los motores de búsqueda actuales, es un mundo complejo y muy avanzado. No obstante, simplificamos este paso reduciéndolo a un filtrado simple de las palabras clave introducidas por el usuario para poder focalizar el esfuerzo de investigación del TFM en la extracción de hechos relevantes de un conjunto de noticias.
- 2) Una vez tenemos el *subset* de noticias relativas a una misma temática, aplicamos **modelado de tópicos** para identificar qué diferentes hilos se han generado relativos a esa temática
- 3) El siguiente paso consistiría en **clasificar qué tópicos son relevantes y cuáles no**. Para ello planteamos el siguiente algoritmo:
 - Agrupamos el *subset* de publicaciones por tópico y mes, contabilizando el total de publicaciones
 - Identificamos el mes que más publicaciones ha tenido cada tópico y calculamos la media
 - Seleccionamos los tópicos que tienen un número de publicaciones mensuales máximo superior a esta media. Éstos serán los tópicos relevantes
- 4) Por último, aplicaremos técnicas de **resumen** o *summarization* para extraer la idea principal del tópico en el mes de más publicaciones esperando que refleje apropiadamente el hecho relevante que ha causado el incremento de publicaciones sobre esa cuestión específica

Capítulo IV.

Estudio del estado del arte

CAPÍTULO IV. ESTUDIO DEL ESTADO DEL ARTE

Podemos por tanto dividir nuestro problema en dos:

- 1) Identificación de tópicos o “temáticas” diferentes que se están comentando con respecto a un argumento, que abordaremos con el modelado de tópicos
- 2) Extracción de qué información es clave o relevante en dichos tópicos, que acometeremos con tareas de resumen

Como ya hemos comentado, ambos problemas son tareas específicas en el ámbito del NLP que llevan años abordándose desde diferentes enfoques y tecnologías. Hacemos un estudio del estado del arte en estas tareas NLP con objeto de identificar cuáles de las actuales tendencias ajustan mejor a nuestra problemática y utilizaremos en el diseño de la solución.

1. IDENTIFICACIÓN DE TÓPICOS. TOPIC MODELLING

Tal y como nos ilustra el artículo de Sagar Pundir en su artículo “New way of topic modelling” de la revista científica “Towards data science” [8], la técnica LDA, *Latent Dirichlet Allocation*, supuso una revolución en el ámbito del *topic modelling*, convirtiéndose en la técnica claramente predominante de los últimos años.

Estudiando las principales técnicas de modelado de tópicos, artículo [9] de la revista “Medium” que analiza las técnicas más populares y data de mayo del 2018, éstas son: LSA (*Latent Semantic Analysis*), pLSA (*Probabilistic Latent Semantic Analysis*) y LDA.

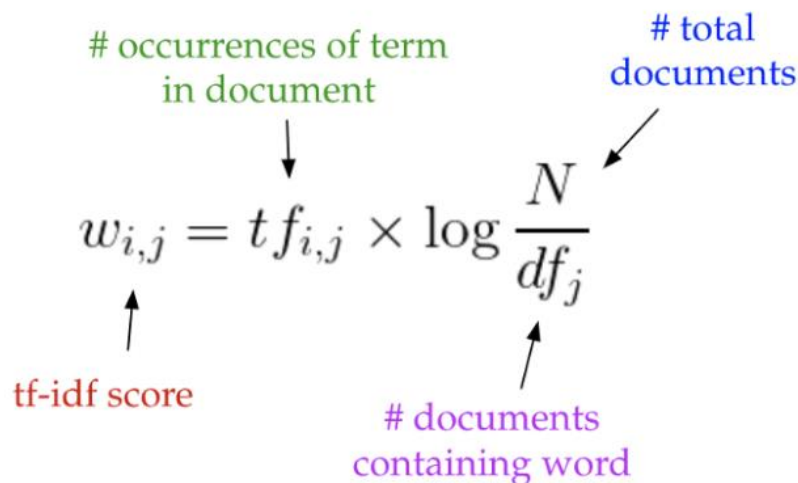
El modelado de tópicos se basa en la idea conceptual de que distintos documentos (o reseñas, *reviews...*) comparten tópicos con distinto peso específico. Es decir, un documento puede tener asociados varios tópicos con diferentes porcentajes de predominancia. Como si habláramos de un tema principal y comentáramos también otros temas. Por otro lado, los tópicos son conjuntos de palabras con, también, un peso específico para cada una de ellas. Por supuesto, una gran mayoría de palabras existe en casi todos los tópicos y lo que cambia es precisamente la distribución de pesos.

LSA:

Fue una de las técnicas pioneras en el modelado de tópicos. Se fundamenta en la construcción de una matriz término-documento que se descompone en dos matrices: documento-tópico y tópico-término.

Para la construcción de las matrices utiliza el score tf-idf que se calcula de la siguiente manera:

Figura 2: Tf-idf o term frequency-inverse document frequency



The diagram shows the formula $w_{i,j} = t f_{i,j} \times \log \frac{N}{d f_j}$ with arrows pointing to its components:

- $t f_{i,j}$ is labeled "# occurrences of term in document" (green) and "tf-idf score" (red).
- N is labeled "# total documents" (blue).
- $d f_j$ is labeled "# documents containing word" (purple).

Fuente: Artículo Tf-idf o term frequency-inverse document frequency de Joyce Xu [9]

La formulación sigue la idea intuitiva de que si una palabra o término aparece muchas veces en un documento debe tener más peso relativo en dicho documento, pero penalizando de manera importante que el término sea también muy frecuente en el resto de los documentos (es decir, palabras frecuentes y poco significativas).

Una vez tenemos construida nuestra matriz documento-término podríamos comenzar a calcular probabilidades para identificar los tópicos latentes que existen en los documentos, pero necesitamos primero reducir la dimensionalidad, ya que tendremos información demasiado dispersa y un alto nivel de ruido. Finalmente, aplicando medidas como *cosine similarity* podremos identificar la similitud entre documentos u términos.

pLSA:

O *Probabilistic Latent Semantic Analysis*, utiliza un método probabilístico en lugar del método de reducción de dimensionalidad. Conceptualmente trata de encontrar un modelo probabilístico que a partir de los tópicos sea capaz de generar los datos observados; es decir, los documentos.

Figura 3: Probabilidad de ver un documento y un término de manera conjunta

$$P(D, W) = \sum_Z P(Z) P(D|Z) P(W|Z)$$

Fuente: Artículo Tf-idf o term frequency-inverse document frequency de Joyce Xu [9]

LDA:

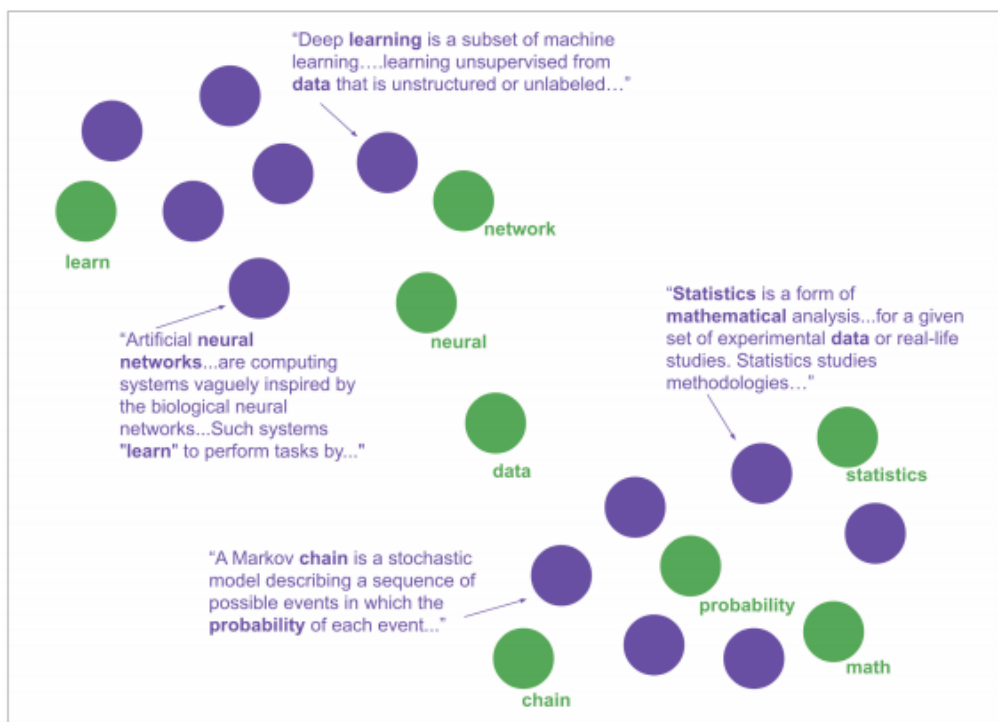
Es la versión Bayesiana de de pLSA. Utiliza como modelo probabilístico la distribución Dirichlet y aportó al ámbito del *topic modelling* capacidad para generalizar bien ante nuevos documentos que sea necesario analizar.

Volviendo al artículo “New way of topic modelling” [8], aunque LDA ha sido el método indiscutible de *topic modelling* durante los últimos años, tiene algunas carencias claras:

- 1) El número de tópicos debe ser conocido
- 2) Necesitamos eliminar *stop-words* (palabras muy habituales con poco significado) para no contaminar los resultados y eliminar ruido
- 3) Trabaja con BOW (*bag of words*), por lo que no tenemos en cuenta el significado semántico de las palabras (mal tratamiento de los sinónimos, homónimos...)

Recientemente están surgiendo nuevos métodos de modelado de tópicos que intentan solventar estos problemas. Este es el caso de **Top2Vec**. Se basa en la creación de un **espacio semántico de documentos y términos representados de manera conjunta**.

Figura 4: Representación del espacio semántico de embeddings de doc. (morado) y palabras (verde)



Fuente: TOP2VEC: DISTRIBUTED REPRESENTATIONS OF TOPICS [10]

Decidimos probar como parte de la solución de este TFM este método.

2. EXTRACCIÓN DE INFORMACIÓN RELEVANTE Y RESÚMEN. SUMMARIZATION

El campo del NLP que se ocupa de extraer información relevante de uno o varios textos y/o realizar un resumen comprensible de los aspectos más importantes del texto se llama *text summarization*. Existen dos enfoques diferentes para abordar este problema:

- 3) **Extractive summarization:** Define un ranking de las frases incluidas en el texto valorando su relevancia. El resumen que obtenemos como salida son las frases top de este ranking. Es decir, es una extracción de frases completas del texto
- 4) **Abstractive summarization:** Utiliza técnicas de *comprensión* y de *generación* del texto para hacer un resumen de la información relevante. El texto que se obtiene como salida no coincidirá literalmente con ningún extracto del contenido analizado

Las últimas tendencias en tareas de *summarization* están claramente lideradas por las tecnologías **Transformers**, creadas en el año 2017 por científicos de Google e investigadores de la Universidad de Toronto. Tomamos como referencia para entender cómo funcionan el sitio oficial de publicación y contenidos de la comunidad Hugging Face [11], en concreto, el curso oficial que han publicado este año 2021 [12]. Hugging Face es una comunidad internacional de expertos en IA que lidera la vanguardia de NLP. Construye, entrena y desarrolla los últimos avances en el ámbito de Transformers, compartiendo e impulsando su uso y desarrollo. Hugging Face tiene en su biblioteca de Transformers más de 7.000 modelos preentrenados en 164 idiomas.

Transformers:

Los Transformers son modelos preentrenados que pueden abordar todo tipo de tareas NLP. Estas arquitecturas de aprendizaje profundo han estado detrás de los modelos más grandes que se han construido recientemente de la mano de grandes tecnológicas como Google, Facebook o Microsoft. A diferencia de las redes neuronales recurrentes, RNN, los Transformers podían paralelizarse de manera muy eficiente, por lo que, con recursos de hardware suficientes podían llegar a entrenar modelos realmente grandes. Entre los más importantes están:

- 5) GPT (junio 2018). Creado por OpenAI
- 6) BERT (octubre 2018). Creado por Google
- 7) GPT-2 (febrero 2019). Creado por OpenAI
- 8) BART y T5 (octubre 2019). Creado por Google
- 9) GPT-3 (mayo 2020). Creado por OpenAI. Se entrenó con más de 45 TB de texto, incluida toda la web pública

Todos los modelos de Transformers han sido entrenados de forma automática como *Language Models*. Esto significa que obtenemos representaciones numéricas de un texto con un avanzadísimo conocimiento estadístico del mismo. Estas representaciones incluyen información semántica y sintáctica. Son capaces de distinguir homónimos (“Mañana tengo que ir al banco”, “Estuve toda la tarde sentada en el banco”), tienen en cuenta el orden de las palabras (“Ana fue a buscar problemas”, “El problema fue a buscar a Ana”), e identifican patrones y estructuras sintácticas.

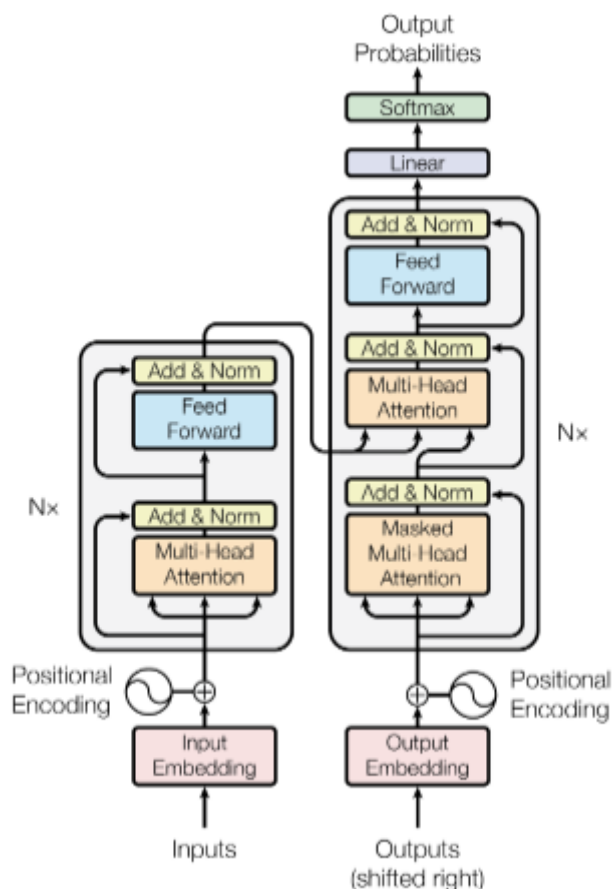
Para crear estos *Languages Models* se han entrenado de manera automática grandes cantidades de texto. Existen dos estrategias de entrenamiento que generan por tanto dos tipos de *Language Models* diferentes:

- 10) *Causal Language Model*: Son capaces de predecir la siguiente palabra. Conocen el contexto previo, pero no futuro
- 11) *Masked Language Model*: Predicen una palabra enmascarada en una frase. Conocen el contexto previo y futuro

Los *Languages Models* tienen mucha información y conocimiento del lenguaje, pero no pueden resolver una tarea NLP específica. Para lograrlo nos bastará con realizar una tarea llamada *fine-tuning*. Partiendo del modelo ya entrenado, volvemos a entrenarlo para la tarea específica que queramos realizar. Un entrenamiento de *fine-tuning* con muy pocos datos consigue resultados sorprendentes.

La arquitectura Transformers es la siguiente:

Figura 5: Arquitectura de los modelos Transformers



Fuente: Hugging Face [12]

La parte de la izquierda de la arquitectura es el *Encoder*, y la de la derecha el *Decoder*. Se pueden utilizar de manera independiente:

- 1) *Encoder-only models*: Para tareas que requieren gran comprensión del *input* como *Sentence Classification*, *Question/Answering* o *NER*, *Named Entity Recognition*. En cada paso la capa de atención puede acceder a todas las palabras de una frase. Es bidireccional. Devuelven una representación vectorial de cada palabra que representa a dicha palabra y su contexto. El tamaño del vector depende de la arquitectura implementada. Ejemplo: BERT
- 2) *Decoder-only models*: Para tareas generativas como *Text Generation*. Funcionan muy parecido a los *encoders* y también nos proporcionan un vector por palabra, pero se diferencian en la capa de atención, que enmascara las palabras a la izquierda o a la derecha de forma que sólo tendremos el contexto de un lado. Estos modelos brillan en tareas causales. Ejemplo: GPT-2

- 3) *Encoder-Decoder* o *Sequence2Sequence models*: Para tareas generativas que requieren un input como *Summatization* o *Translation*. En el caso por ejemplo de una traducción del inglés al español, el *encoder* captaría muy bien el contexto y significado de una palabra en inglés, y el *decoder* generaría la traducción en español, que tiene una estructura completamente diferente (orden, número de palabras, normas sintácticas...). Ejemplo: T5

Los aspectos más innovadores de esta arquitectura son:

- 1) *Positional Encoding*: Registra información del orden de las palabras. Esto no había existido en las arquitecturas previas a Transformers
- 2) *Attention Layer*: Es una capa que nos dice para cada palabra a qué palabras tenemos que prestar atención y cuáles podemos ignorar

Para nuestro Generador de Timelines probaré distintos tipos de modelos Transformers compartidos en Hugging Face. Pero nos tenemos que limitar a aquellos que han sido entrenados en español. Valoraré según los resultados que vaya obteniendo si es necesario hacer *fine-tunning* de algún modelo para lograr un mejor ajuste a la tarea que quiero realizar.

Un problema que anticipo en el uso de Transformers para tareas de resumen o *summarization* es que los Transformers tienen limitada la entrada a un número máximo de palabras o símbolos (tokens). Para salvar esta limitación probaré a combinar *Extractive Summarization*, con objeto de obtener las frases más relevantes de cada noticia, y *Abstractive Summarization* a partir de ellas.

- 1) *Extractive Summarization*. La abordaré con Sentence Transformers [11]
- 2) *Abstractive Summarization*. La abordaré con modelos entrenados específicamente para tareas de resumen Transformers [11]

Capítulo V. Análisis de alternativas topic modelling

CAPÍTULO V. ANÁLISIS DE ALTERNATIVAS TOPIC MODELLING

3. FUNCIONAMIENTO TOP2VEC

Top2Vec es un modelo muy reciente, año 2020, que aborda el modelado de tópicos desde una perspectiva diferente a las de técnicas previas. Según se explica en su *paper* [10], basa la identificación de tópicos en la similitud semántica que existe entre los documentos. Para ello, genera como primer paso las representaciones numéricas (*embeddings*) de palabras y documentos, incrustándolas en un espacio vectorial semántico. Una vez que los documentos y las palabras están representadas en ese espacio vectorial, el algoritmo busca grupos densos de documentos, que serán los tópicos, y las palabras que atrajeron a los documentos a esas áreas densas serán las palabras que conforman el tópico. El algoritmo está estructurado de la siguiente forma:

- 1) Creación de las representaciones numéricas de palabras y documentos y representación conjunta de ambos en el espacio vectorial. Es paso puede realizarse utilizando distintos modelos de *embeddings*
 - Doc2vec. Entrena un modelo con doc2vec desde cero
 - Universal Sentence Encoder. Modelos pre-entrenados
 - Bert Sentence transformer. Modelos pre-entrenados
- 2) Reducción dimensional con UMAP. Las representaciones vectoriales de documentos en espacios de dimensiones altas están muy dispersos y es necesario reducir la dimensionalidad para poder *clusterizarlos*
- 3) Identificación de áreas densas con HDBSCAN
- 4) Cálculo del centroide (o nodo central) para cada área densa en el espacio dimensional original. Ese será el tópico
- 5) Búsqueda de los n vectores de palabras más cercanos a cada centroide según proximidad semántica

El parámetro fundamental que tenemos que definir para la creación de nuestro modelo con Top2Vec, y que influirá de manera significativa en los resultados, es el modelo de *embedding* que elijamos. Decido analizar el funcionamiento de doc2vec y BERT Sentence Transformer, que es un modelo pre-entrenado por Sentence Transformer que ha proporcionado resultados muy buenos en la generación de tópicos y tareas de cálculo de similitud semántica.

4. ALTERNATIVA 1: TOP2VEC CON BERT SENTENCE TRANSFORMER

Comenzamos analizando cómo funciona Top2Vec cuando realiza el *embedding* con el modelo “distiluse-base-multilingual-cased”, que es un modelo pre-entrenado a partir del modelo BERT por Sentence Transformer. Al utilizar un modelo pre-entrenado, los tiempos de procesamiento para la generación de los tópicos serán mucho más cortos y esperamos que los resultados sean bastante buenos.

Comenzamos probando con la temática “pandemia”. Queremos generar un timeline que nos explique qué hechos relevantes han ocurrido relativos a este término desde el 01/01/2020 hasta el 31/05/2021. Para ello, como ya comenté en el diseño de la solución que planteé en el apartado 3.2 de este documento, vamos a simplificar esta parte limitándonos a un filtrado simple de las publicaciones que contienen este término:

```
import gensim
from top2vec import Top2Vec
```

```
#Preparamos el contenido en una lista para aplicar top2vec
content = [content for content in news_data['contenido'] if 'pandemia' in content.lower()]
content[100]
```

'A mediados de junio de 2020, el Congreso aprobó el ingreso mínimo vital sin ni un solo voto en contra. Se trataba de una ayuda que ya entraba en los planes del Gobierno para un futuro algo más lejano de lo que fue, ya que la crisis provocada por la pandemia aceleró su tramitación. Así, salía adelante la prestación de mayor calado social, enfocada a la población sin ingresos y más vulnerable. La altísima demanda de solicitudes generó un cuello de botella que llevó al Ejecutivo a modificar en varias ocasiones el decreto-ley para introducir cambios y nuevos requisitos con el objetivo de que llegase a más gente. Pese a todo, casi un año después... el IMV solo está llegando a una de cada cuatro familias de las estimadas inicialmente. Con todo, son muchos los que todavía siguen esperando la llegada del ingreso mínimo vital. Aquellos que están en esta situación tienen una opción

```
len(content)
```

```
17205
```

Obtenemos un *subset* de 17.205 publicaciones. Todo lo ocurrido con respecto a la “pandemia” ha sido sin duda lo más destacado en el periodo de fechas de mi conjunto de datos.

Procedemos a generar el modelo de tópicos con Top2Vec. Elegimos la “velocidad” más lenta, “deep-learn”, frente a las opciones más rápidas “fast-learn” o “learn” porque consigue resultados bastante superiores.

```
model = Top2Vec(documents=content, embedding_model="distiluse-base-multilingual-cased", speed="deep-learn", workers=4)
```

```
2021-09-03 12:44:18,941 - top2vec - INFO - Pre-processing documents for training
2021-09-03 12:44:43,401 - top2vec - INFO - Downloading distiluse-base-multilingual-cased model
2021-09-03 12:44:45,501 - top2vec - INFO - Creating joint document/word embedding
2021-09-03 13:12:30,190 - top2vec - INFO - Creating lower dimension embedding of documents
2021-09-03 13:12:55,524 - top2vec - INFO - Finding dense areas of documents
2021-09-03 13:12:58,120 - top2vec - INFO - Finding topics
```

```
model.get_num_topics()
```

```
91
```

Hemos tardado 29 minutos en generar el modelo y nos devuelve 91 tópicos.

El modelo generado tiene dos conjuntos de vectores:

- Tópicos. O relación de las palabras con los tópicos. Son listas de palabras frecuentes de cada hilo o tópico generado con el peso relativo que tiene dicha palabra en ese tópico.

- Relación de documentos con los tópicos. Son listas de documentos con el peso relativo de los documentos en ese tópico. También con una relación *many-to-many*.

[illegible]

A word cloud on a black background featuring various sports-related terms. The most prominent words are 'fútbol' (green), 'olimpicos' (light green), 'torneo' (teal), and 'olimpico' (yellow). Other visible words include 'deportivo', 'deportiva', 'sports', 'atletas', 'gimnasios', 'deporte', 'entrenador', 'league', 'deportivas', 'deportivos', 'madrid', 'equipos', 'competitividad', 'deportistas', 'competicion', 'competitivo', 'competiciones', 'antivirales', 'tenis', 'gimnasio', 'medalla', 'deportes', 'olimpica', 'epidemiologicos', 'ebola', 'zidane', 'equipo', 'espanoles', 'campeonato', 'torneos', 'atletico', 'futbolistas', 'club', 'concurso', 'champions', 'baloncesto', 'deportista', 'vacunaciones', 'competencia', 'epidemiologicos', 'clubes', 'tenista', 'coronavirus', 'eur copa', and 'antivirales'.

[illegible]

A word cloud of Spanish terms related to vaccination. The most prominent words are 'vacunaciones', 'vacuna', 'vacunada', 'vacunadas', 'vacunando', 'vacunacion', 'vacunarse', and 'vacunados'. Other visible words include 'epidemia', 'vacuna', 'ebola', 'gripe', 'coronavirus', 'infecciones', 'infectados', 'infectar', 'infectadas', 'infectarse', 'infectado', 'infectad', 'enfermedad', 'enfermos', 'enfermedades', 'sanitarias', 'espanola', 'espana', 'espanoles', 'contagios', 'contagioso', 'epidemiologico', 'epidemiologia', 'epidemiologicos', 'antivirales', 'sanidad', 'contag', 'viraes', 'virus', 'epidemiologicas', 'antiviral', 'infectada', 'contagiosa', 'infeccion', 'adenovirus', 'coronavirus', 'epidemiologica', and 'infectadas'. The words are arranged in a dense, overlapping manner with varying font sizes and colors (primarily shades of green, blue, and purple).

[illegible][illegible]

A word cloud of Spanish terms related to healthcare and medicine. The most prominent words are 'sanitarios' (green), 'sanitaria' (yellow), 'sanidad' (yellow), 'sanitario' (purple), and 'hospitalizados' (purple). Other visible words include 'enfermos', 'epidemiológicos', 'medica', 'enfermero', 'epidemiologo', 'hospitalizadas', 'enferma', 'doctores', 'salud', 'health', 'vacuna', 'hospitalizacion', 'vacunando', 'clinic', 'vacunas', 'ambulancia', 'vacunaciones', 'vacunados', 'clinica', 'vacunada', 'medical', 'hospital', 'clínicas', 'diagnosticaron', 'doctora', 'hospitalizado', 'epidemiólogos', 'epidemiología', 'vacunarse', 'vacunacion', 'enfermeras', 'medicos', 'medico', 'epidemiologica', 'vacunadas', 'enfermeros', and 'hospitalales'. The words are arranged in a dense, overlapping manner with varying font sizes and colors.

[illegible][illegible][illegible]

Probamos con otros *subset* para validar el funcionamiento en un modo más amplio. Queremos probar, además de otras temáticas y por tanto, contextos diferentes, *subsets* que difieran bastante en número total de publicaciones. Seleccionamos por este último criterio los siguientes:

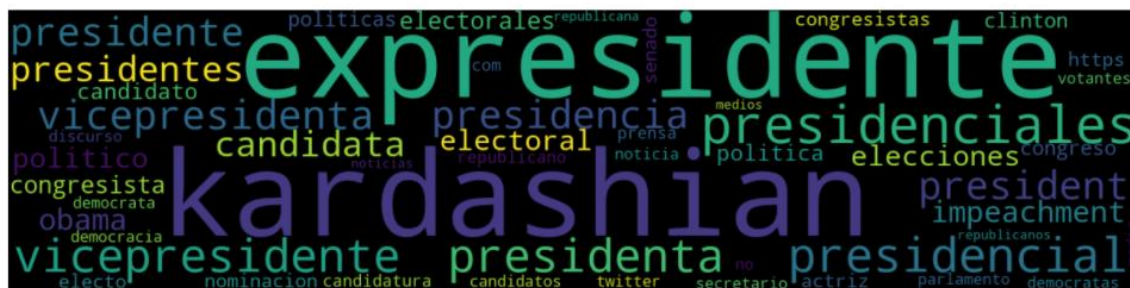
- Trump
- China
- 5G
- Elecciones

Tabla 1: Resultados modelado de tópicos Top2Vec con BERT Sentence Transformer

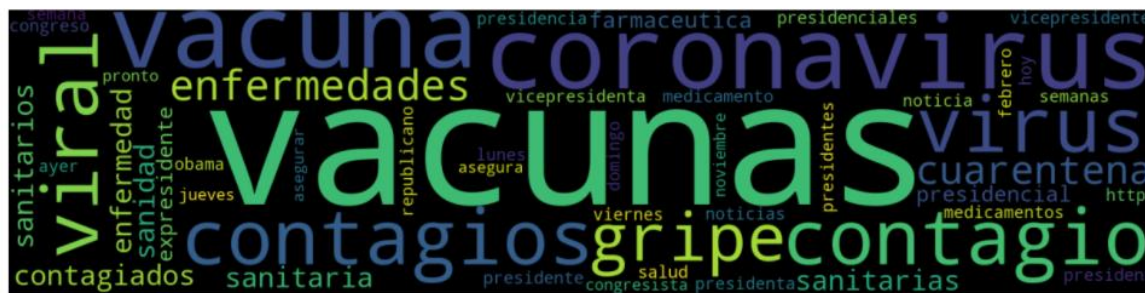
TEMÁTICA	TOTAL PUBLICACIONES	Nº TÓPICOS	TIEMPO DE EJECUCIÓN (min.)
Pandemia	17.205	91	20
Trump	1.920	15	4
China	3.167	2	6
5G	542	7	2
Elecciones	3.955	3	

Cuando analizamos temáticas con pocas publicaciones el modelo BERT de Sentence Transformer vemos que funciona aceptablemente bien en algunos casos, como el caso de “Trump” y “5G”, con 15 y 7 tópicos respectivamente. Pero no parece funcionar tan bien en el caso de “China” y “Elecciones”. Analizando las nubes de palabras de **“Trump”** comprobamos que cubre temáticas concretas amplias:

Topic 0



Topic 1



[illegible]

A word cloud featuring various political terms. The most prominent words are 'presidenciales' (green), 'senado' (yellow), and 'republicano' (blue). Other visible words include 'democrata', 'senadores', 'política', 'senador', 'políticas', 'electoral', 'impeachment', 'gubernadores', 'semana', 'elecciones', 'president', 'legisladores', 'congresistas', 'vicepresidente', 'protesta', 'republicana', 'constitución', 'congreso', 'presidentes', 'protestas', 'votantes', 'parlamento', 'vicepresidenta', 'expresidente', 'manifestantes', 'viernes', 'político', 'presidencia', 'federales', 'domingo', 'ayer', 'febrero', 'obama', 'manifiestantes', 'lunes', 'jueves', 'semanas', 'electoral', 'domingo', 'protesta', 'republicana', 'constitución', 'congreso', 'presidentes', 'protestas', 'votantes', 'parlamento', 'vicepresidenta', 'expresidente', 'manifestantes', 'viernes', 'político', 'presidencia', 'federales', 'domingo', 'ayer', 'febrero', 'obama', 'manifiestantes', 'lunes', 'jueves', 'semanas', 'electoral'.

[illegible][illegible]

A word cloud featuring various terms related to the US-Mexico border and immigration. The most prominent words are 'inmigración', 'mexico', 'inmigrantes', and 'embajada'. Other visible words include 'fronteras', 'presidentes', 'expresidente', 'presidencia', 'extranjeros', 'impeachment', 'españoles', 'democratas', 'obama', 'america', 'presidenta', 'republicana', 'vicepresidente', 'congresista', 'nacional', 'muro', 'ayer', 'viernes', 'ciudadano', 'febrero', 'ciudadanos', 'presidencial', 'terminos', 'restricciones', 'republicano', 'estadounidenses', 'federal', 'congresistas', 'president', 'federales', 'presidentes', 'nacionales', 'estadounidense', 'lunes', 'presidenciales', 'americano', 'espana', 'semana', 'texas', 'venezuela', 'racismo', 'norteamericano', 'ciudadanos', 'vicepresidenta', 'republicanos', and 'impeachment'. The words are in various colors (green, blue, purple, yellow) and sizes, creating a dense visual representation of the topic.

Topic 0



A word cloud featuring various terms related to elections. The most prominent words are 'election', 'parlamento', 'electoral', 'parlamentaria', 'electorales', 'parlamentario', 'elector', 'politico', 'congreso', 'legislatura', 'parlamentarias', 'electores', 'elegida', 'parlamentarios', 'referendum', 'presidencial', 'democratica', 'democracia', 'electo', 'politico', 'elegido', 'politico', 'voto', 'votada', 'presidenciales', 'vicepresidente', 'votaron', 'eleccion', 'reeleccion', 'candidato', 'votacion', 'votante', 'votaciones', 'reelegido', 'parlamentaria', 'votado', 'vota', 'democratica', 'democracia', 'electo', 'politico', 'elegido', 'politico', 'voto', 'votada', 'presidenciales', 'vicepresidente', 'votaron', 'eleccion', 'reeleccion', 'candidato', 'votacion', 'votante', 'votaciones', 'reelegido', 'parlamentaria', 'votado', 'vota', 'democratica', 'democracia', 'electo', 'politico', 'elegido', 'politico', 'voto', 'votada', 'presidenciales', 'vicepresidente'. Other visible words include 'votar', 'opciones', 'democratico', 'netanyahu', 'democraticas', 'legislatura', 'elegidos', 'votantes', 'votaran', 'parlamentaria', 'electoral', 'reelegido', 'parlamentario', 'electorales', 'parlamentario', 'elector', 'politico', 'congreso', 'legislatura', 'parlamentarias', 'electores', 'elegida', 'parlamentarios', 'referendum', 'presidencial', 'democratica', 'democracia', 'electo', 'politico', 'elegido', 'politico', 'voto', 'votada', 'presidenciales', 'vicepresidente', 'votaron', 'eleccion', 'reeleccion', 'candidato', 'votacion', 'votante', 'votaciones', 'reelegido', 'parlamentaria', 'votado', 'vota', 'democratica', 'democracia', 'electo', 'politico', 'elegido', 'politico', 'voto', 'votada', 'presidenciales', 'vicepresidente'.

No se observa una correlación sobre cuándo funciona mejor o peor el modelo y el número total de publicaciones. Entendemos que está más bien relacionado con lo “diferentes” que sean los tópicos a identificar. A pesar de que sabemos que el tiempo de computación para la generación del modelo va a ser muy superior, decidimos analizar el funcionamiento de Top2Vec con doc2vec para el proceso de *embedding*.

5. ALTERNATIVA 2: TOP2VEC CON DOC2VEC

Si utilizamos doc2vec para el proceso de *embedding*, no estamos utilizando modelos pre-entrenados si no que entrenamos uno desde cero. Los resultados deberían de ser mucho más afinados, pero con tiempos de computación elevadísimos. Probamos las mismas opciones del apartado anterior para poder hacer una comparativa.

Comenzando con el *subset* con más número de publicaciones, “pandemia”, obtenemos un tiempo de procesamiento de 3 horas y 42 minutos.

```
model = Top2Vec(documents=content, speed="deep-learn", workers=4)
2021-09-18 18:55:25,731 - top2vec - INFO - Pre-processing documents for training
2021-09-18 18:55:47,050 - top2vec - INFO - Creating joint document/word embedding
2021-09-18 22:34:50,703 - top2vec - INFO - Creating lower dimension embedding of documents
2021-09-18 22:35:16,293 - top2vec - INFO - Finding dense areas of documents
2021-09-18 22:35:21,032 - top2vec - INFO - Finding topics
```

Este tiempo no sería aceptable desde el punto de vista de implementación de una aplicación on-line o app para la generación de timelines. Es cierto que los tiempos que estoy obteniendo son con CPU y con un equipo hardware bastante básico, pero, en principio, no es aceptable si lo comparamos con los tiempos que obtuve con BERT, 20 minutos para este caso, el más voluminoso.

Analizamos todos los casos y todos los parámetros del apartado anterior. Los resultados son:

Tabla 2: Resultados modelado de tópicos con Top2Vec con doc2vec

TEMÁTICA	TOTAL PUBLICACIONES	Nº TÓPICOS	TIEMPO DE EJECUCIÓN (min.)
Pandemia	17.205	180	222
Trump	1.920	21	8
China	3.167	2	21
5G	542	5	2
Elecciones	3.955	53	27

Fuente: Elaboración propia

La temática de “China” sigue sin arrojar mejores resultados y las temáticas “5G” y “Trump” obtienen resultados muy similares. Sin embargo, la temática “elecciones” sí mejora de manera muy significativa con unos tiempos de ejecución más o menos aceptables considerando que tenemos unas condiciones de hardware mínimas.

Sí hay que descartar absolutamente para el caso de publicaciones muy altas como ocurre con “pandemia” esta opción, ya que los tiempos de ejecución no son asumibles y los resultados en cuanto a tópicos identificados se disparan. Por mucho que definamos criterios a posteriori de relevancia, parece demasiado difuso para el objetivo que nos atañe, que es el de identificar hechos relevantes concretos.

Capítulo V. Análisis de alternativas summarization

CAPÍTULO V. ANÁLISIS DE ALTERNATIVAS

SUMMARIZATION

Con los resultados obtenidos en el capítulo anterior, “Análisis de alternativas *topic modelling*” decidimos avanzar en el análisis de alternativas de *summarization* o extracción de información relevante con las temáticas “pandemia” y “Trump”, con el objetivo de analizar el funcionamiento de la herramienta que estamos construyendo con temáticas muy voluminosas y extensas, pero también con otras más específicas y concretas.

Antes de abordar las tareas de *summarization* debemos identificar qué tópicos son relevantes y cuáles no. Para ello realizamos los siguientes pasos:

- **Construimos un *dataset*** que incluya tanto la información relativa a las publicaciones (contenido, fecha...) como los **tópicos asociados** a cada publicación y el **score** o peso relativo de dicho tópico en el documento

```
#Creamos listas con la relación num_documento, topic, score
num_topics=model.get_num_topics()
documents_by_topic=[]
for topic in range(1, num_topics):
    documents, document_scores, document_ids = model.search_documents_by_topic(topic_num=topic, num_docs=20)
    a_topic=[(doc_id, topic, score) for (score, doc_id) in zip(document_scores, document_ids)]
    documents_by_topic.append(a_topic)
documents_by_topic
[(9586, 3, 0.90624774),
 (16516, 3, 0.90455127),
 (10674, 3, 0.90294206),
 (16833, 3, 0.89982045),
 (7020, 3, 0.8989577),
 (14309, 3, 0.8980471),
 (8451, 3, 0.8968386),
 (13525, 3, 0.89678566)]
```

En este punto debemos indicar el número de documentos por tópico que se cargará. 20 nos parece un número razonable. Serán los 20 documentos más relevantes para un tópico dado.

Continuamos construyendo el *dataset*.

```
#Hacemos flatten porque no necesitamos tener la lista en dos niveles
documents_by_topic = [val for sublist in documents_by_topic for val in sublist]
documents_by_topic
[(15377, 1, 0.7749876),
 (15676, 1, 0.76874536),
 (16964, 1, 0.7682978),
 (13019, 1, 0.7679247),
 (13487, 1, 0.7640743),
 (16962, 1, 0.75708306),
 (15820, 1, 0.7526659),
 (1107, 1, 0.7507368)]

content_filter = [(id, new) for (id, new) in enumerate(news_data['contenido']) if 'pandemia' in new.lower()]
id, new = zip(*content_filter)
#Obtenemos un subset de news_data
news_data_filter = pd.DataFrame([news_data.iloc[i] for i in id], )
news_data_filter[:10]
```

	id	date	username	nlikes	url	titulo	contenido	fecha
17	1399101271647342594	2021-05-30 22:31:51	20m	4	https://t.co/0UHgZQXYIZ	Estafas mutantes: la pandemia hace que los del...	Los expertos en ciberseguridad están alertando...	30.05.2021 - 22:08h
28	1399084095280488452	2021-05-30 21:23:36	20m	3	https://t.co/mlVWneQ4I3	El zasca de Rosa Benito a Carmen Borrego por s...	Rosa Benito fue una de las personas que se uni...	30.05.2021 - 21:14h
54	1399045050220235248	2021-05-30	20m	10	https://t.co/0EeBPC0K4	"No puedes definir 32 putos	Cada cierto tiempo, dentro de	30.05.2021 -

Generación automática de timelines con Transformers

```
#Necesitamos regenerar el índice para enlazar con el dataframe "documents_by_topic"
news_data_filter=news_data_filter.reset_index()
news_data_filter.head()
```

```
#Lo enlazamos con documents_by_topic
news_data_filter['id_document']=news_data_filter.index
documents_by_topic=pd.DataFrame(documents_by_topic, columns=('id_document', 'topic', 'score'))
news_data_filter_topics=pd.merge(documents_by_topic, news_data_filter, on='id_document')
news_data_filter_topics.head()
```

	id_document	topic	score	index	id	date	username	likes	url	titulo	contenido	fecha
0	15377	1	0.774988	77069	1251043964993634305	2020-04-17 07:04:59	20m	2	https://t.co/Cu5bPCgLDX	El Atlético jugará la próxima edición de Champ...	Buenas noticias en el Wanda Metropolitano. Con...	17.04.2020 - 08:12h
1	15676	1	0.768745	78195	1248657310190272512	2020-04-10 17:01:16	20m	4	https://t.co/y8XNth8VCq	Vinicius y Joao Félix juegan por Madrid y Atlé...	El mundo del fútbol continúa paralizado por la...	10.04.2020 - 18:44h
2	16964	1	0.768298	83055	1239960604217085954	2020-03-17 17:03:40	20m	13	https://t.co/lQ2o7K29I7	Deportistas e influencers se retan en un	La crisis sanitaria por la pandemia de	17.03.2020 - 14:51h

Obtenemos un *dataframe* de 1800 registros.

```
len(news_data_filter_topics)
```

1800

```
topics_bert=pd.DataFrame(news_data_filter_topics, columns=('id_document', 'topic', 'score', 'titulo', 'contenido', 'date'))
```

- El siguiente paso será agrupar las publicaciones por tópico, año y mes, seleccionando los tópicos que tengan un número de publicaciones mensuales superior a la media en algún mes. Es decir, **identificamos los tópicos relevantes**.

```
: topics_month=topics_bert['id_document']
topics_month.index=pd.to_datetime(topics_bert['date'])
monthly = topics_month.resample('M').count()
monthly
```

```
: date
2020-01-31    1
2020-02-29    0
2020-03-31   121
2020-04-30   189
2020-05-31   182
2020-06-30   109
2020-07-31   137
2020-08-31    69
2020-09-30    99
2020-10-31   113
2020-11-30   126
2020-12-31   139
2021-01-31   106
2021-02-28    95
2021-03-31   126
2021-04-30   110
2021-05-31    78
Freq: M, Name: id_document, dtype: int64
```



```
#Generamos dos columnas con el mes y el año a partir de la fecha
topics_bert['fecha']=pd.to_datetime(topics_bert['date'])
topics_bert['mes']=topics_bert['fecha'].dt.month
topics_bert['año']=topics_bert['fecha'].dt.year
```

```
topics_bert.head()
```

	id_document	topic	score	título	contenido	date	fecha	mes	año
0	15377	1	0.774988	El Atlético jugará la próxima edición de Champ...	Buenas noticias en el Wanda Metropolitano. Con...	2020-04-17 07:04:59	2020-04-17 07:04:59	4	2020
1	15676	1	0.768745	Vinicius y Joao Félix juegan por Madrid y Atlé...	El mundo del fútbol continúa paralizado por la...	2020-04-10 17:01:16	2020-04-10 17:01:16	4	2020
2	16964	1	0.768298	Deportistas e influencers se retan en un torne...	La crisis sanitaria por la pandemia de coronav...	2020-03-17 17:03:40	2020-03-17 17:03:40	3	2020
3	13019	1	0.767925	El fútbol español guardará un minuto de silenc...	El fútbol español se solidarizará con todas la...	2020-06-07 12:38:59	2020-06-07 12:38:59	6	2020
4	13487	1	0.764074	El CSD no descarta el fútbol con público al in...	Irene Lozano, presidenta del Consejo Superior ...	2020-05-27 00:37:01	2020-05-27 00:37:01	5	2020

```
#Agrupamos por tópico y mes, contabilizando el número de noticias publicadas. Calculamos por tópico la media y la desviación estándar
temas_por_mes=topics_bert.groupby(by=['año', 'mes', 'topic'], as_index=False)['id_document'].count()
temas_por_mes
```

	año	mes	topic	id_document
0	2020	1	18	1
1	2020	3	1	5
2	2020	3	3	3
3	2020	3	4	1
4	2020	3	6	3
...
799	2021	5	81	3
800	2021	5	83	1
801	2021	5	86	4
802	2021	5	88	1
803	2021	5	90	1

804 rows x 4 columns

```
import numpy as np
#Creamos un dataframe año-mes-topic completo para no desvirtuar las estadísticas que obtenga a posteriori
a_list=np.array([], dtype=int)
m_list=np.array([], dtype=int)
t_list=np.array([], dtype=int)
for i in (2020, 2021):
    for j in range(1,13):
        for k in range(1,80):
            t_list=np.append(t_list, k)
            m_list=np.append(m_list, j)
            a_list=np.append(a_list, i)
```

Generación automática de timelines con Transformers

```
#Actualizamos el número de noticias por topico y mes
topicos_per_month_st=pd.merge(d, topicos_per_month, how='outer')
topicos_per_month_st
```

	año	mes	topic	id_document
0	2020	1	1	NaN
1	2020	1	2	NaN
2	2020	1	3	NaN
3	2020	1	4	NaN
4	2020	1	5	NaN
...
1997	2021	5	81	3.0
1998	2021	5	83	1.0
1999	2021	5	86	4.0
2000	2021	5	88	1.0
2001	2021	5	90	1.0

2002 rows x 4 columns

```
#Reemplazamos Los valores NaN por cero y eliminamos los registros a partir de mayo 2021
topicos_per_month_st=topicos_per_month_st.fillna(0)
topicos_per_month_st=topicos_per_month_st.drop(topicos_per_month_st[(topicos_per_month_st['año']==2021) & (topicos_per_month_st['
```

	año	mes	topic	id_document
0	2020	1	1	0.0
1	2020	1	2	0.0
2	2020	1	3	0.0
3	2020	1	4	0.0
4	2020	1	5	0.0
...
1997	2021	5	81	3.0
1998	2021	5	83	1.0
1999	2021	5	86	4.0
2000	2021	5	88	1.0
2001	2021	5	90	1.0

1449 rows x 4 columns

```
#Obtenemos Las estadísticas mensualizadas
topicos_relevantes=topicos_per_month_st.groupby('topic')['id_document'].agg([np.std, np.max, np.min])
topicos_relevantes=topicos_relevantes.round(1)
topicos_relevantes
```

	std	amax	amin
topic			
1	2.1	6.0	0.0
2	1.4	5.0	0.0
3	1.3	4.0	0.0
4	1.4	4.0	0.0
5	1.4	5.0	0.0
...
86	1.0	4.0	1.0
87	1.1	4.0	1.0
88	1.5	5.0	1.0
89	1.6	6.0	1.0
90	1.0	4.0	1.0

90 rows x 3 columns

Generación automática de timelines con Transformers

```
#Seleccionamos los topicos relevantes con el siguiente criterio: el máximo de noticias/mensuales es igual o superior al valor
#máximo medio
mean=int(topicos_relevantes['amax'].mean())
topicos_relevantes_selected=topicos_relevantes[topicos_relevantes['amax']>=mean]
topicos_relevantes_selected
```

	std	amax	amin
topic			
1	2.1	6.0	0.0
2	1.4	5.0	0.0
5	1.4	5.0	0.0
7	2.3	9.0	0.0
8	3.0	10.0	0.0
9	1.6	5.0	0.0
11	1.6	5.0	0.0
13	1.5	5.0	0.0
14	1.6	6.0	0.0
15	2.2	9.0	0.0
16	2.5	10.0	0.0
17	1.8	7.0	0.0
18	1.2	5.0	0.0

Ya tenemos en un único *dataset* toda la información que necesitamos sobre las publicaciones, tanto los contenidos y fecha como los tópicos que tiene asociados:

```
topics_bert.head()
```

	id_document	topic	score	titulo	contenido	date	fecha	mes	año
0	15377	1	0.774988	El Atlético jugará la próxima edición de Champ...	Buenas noticias en el Wanda Metropolitano. Con...	2020-04-17 07:04:59	2020-04-17 07:04:59	4	2020
1	15676	1	0.768745	Vinicius y Joao Félix juegan por Madrid y Atlé...	El mundo del fútbol continúa paralizado por la...	2020-04-10 17:01:16	2020-04-10 17:01:16	4	2020
2	16964	1	0.768298	Deportistas e influencers se retan en un torne...	La crisis sanitaria por la pandemia de coronav...	2020-03-17 17:03:40	2020-03-17 17:03:40	3	2020
3	13019	1	0.767925	El fútbol español guardará un minuto de silenc...	El fútbol español se solidarizará con todas la...	2020-06-07 12:38:59	2020-06-07 12:38:59	6	2020
4	13487	1	0.764074	El CSD no descarta el fútbol con público al in...	Irene Lozano, presidenta del Consejo Superior ...	2020-05-27 00:37:01	2020-05-27 00:37:01	5	2020

Y además qué tópicos son relevantes con el mes que más publicaciones ha tenido.

```
topicos_per_month
```

	año	mes	topic	id_document
0	2020	1	18	1
1	2020	3	1	5
2	2020	3	3	3
3	2020	3	4	1
4	2020	3	6	3
...
799	2021	5	81	3
800	2021	5	83	1
801	2021	5	86	4
802	2021	5	88	1
803	2021	5	90	1

Podemos por tanto abordar las **tareas de summarization** o extracción de información relevante.

Tal y como identificamos en el capítulo IV, “Estudio del estado del arte”, existen dos enfoques diferentes a la hora de abordar las tareas de *summarization*:

- *Extractive summarization*. Extracción literal de frases del texto según un ranking de cuales son más relevantes.
- *Abstractive summarization*. Proceso de comprensión del texto y posterior generación de resumen. No coincidirá literalmente con ninguna parte del texto.

Abordamos ambos enfoques desde modelos pre-entrenados de Transformers.

Extractive summarization:

Utilizaremos los modelos entrenados por Sentence Transformers [13], que tienen aplicaciones específicas para tareas de *extractive summarization*. Este modelo divide el texto en frases y las transforma en vectores numéricos que contienen información semántica de la frase y de su contexto, proceso de *embeddings*. Posteriormente calcula la similitud entre frases aplicando la métrica *cosine similarity*. Finalmente, aplicando el algoritmo *LexRank*, calcula las frases principales de cada tópico.

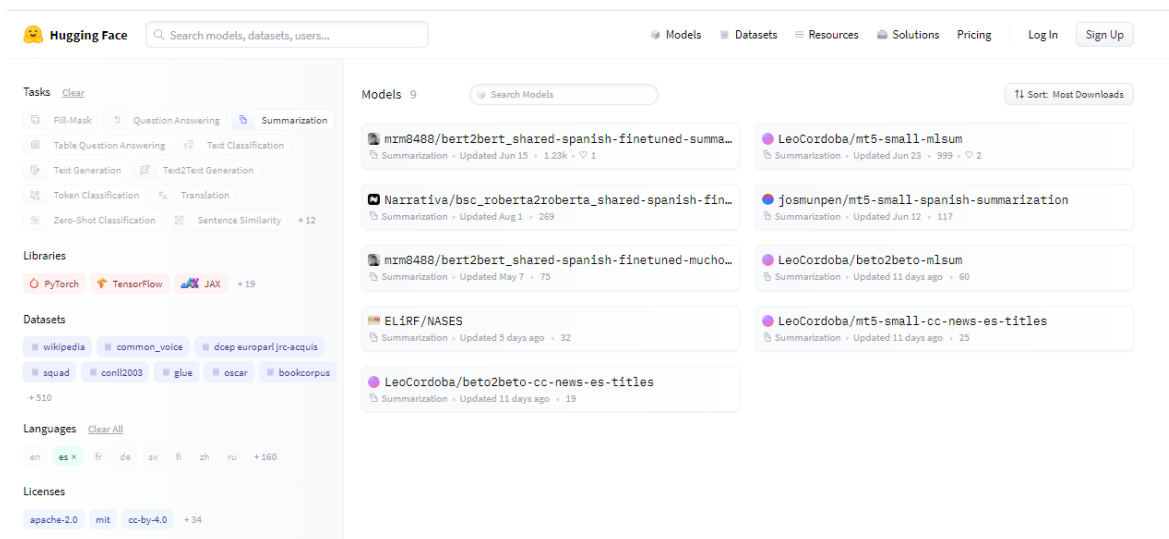
```
os.chdir(r'C:\Python\sentence-transformers-master\sentence-transformers-master\examples\applications\text-summarization')
import LexRank
```

```
import LexRank
import nltk
from sentence_transformers import SentenceTransformer, util
import numpy as np
from LexRank import degree centrality_scores
```

```
model_sentence_transformer = SentenceTransformer('sentence-transformers/all-MiniLM-L6-v2')
```

Abstractive summarization:

Utilizaremos modelos pre-entrenados de la biblioteca de modelos Transformers publicados por Hugging face [11]. Si accedemos a la página donde se alojan los modelos, y filtramos por modelos pre-entrenados para tareas de *summarization* en idioma español obtenemos los siguientes:



Tras algunas pruebas preliminares, optamos por probar con el modelo

“bert2bert_shared-spanish-finetuned-summarization” [14], que es un modelo específicamente pre-entrenado para tareas de *summarization* con un *dataset* de publicaciones de periódicos on-line en 5 lenguas diferentes, francés, alemán, español, ruso y turco.

```
import torch
from transformers import BertTokenizerFast, EncoderDecoderModel
device = 'cuda' if torch.cuda.is_available() else 'cpu'
ckpt = 'mrm8488/bert2bert_shared-spanish-finetuned-summarization'
tokenizer = BertTokenizerFast.from_pretrained(ckpt)
model = EncoderDecoderModel.from_pretrained(ckpt)

The following encoder weights were not tied to the decoder ['bert/pooler']
The following encoder weights were not tied to the decoder ['bert/pooler']
```

Y construimos una función, *generate_summary()*, que nos permite generar un resumen abstractivo utilizando este modelo.

```
def generate_summary(text):
    inputs = tokenizer([text], padding="max_length", truncation=True, max_length=512, return_tensors="pt")
    input_ids = inputs.input_ids.to(device)
    attention_mask = inputs.attention_mask.to(device)
    output = model.generate(input_ids, attention_mask=attention_mask)
    return tokenizer.decode(output[0], skip_special_tokens=True)
```

Respecto a las limitaciones de los modelos, tanto en el caso de Sentence Transformers como del modelo Bert2Bert, tenemos que considerar que las entradas a resumir admiten un máximo de 512 tokens (palabras y símbolos), y son truncadas si superan esta longitud máxima.

Una vez preparados los *datasets* de datos y elegidos los métodos o modelos que vamos a utilizar para realizar la extracción de información relevante **definimos y analizamos la estrategia a implementar. Lo hacemos desde cuatro enfoques diferentes, cuatro alternativas de diseño**, que exponemos a continuación.

1. ALTERNATIVA 1: CONCATENACIÓN DE NOTICIAS Y RESUMEN ABSTRACTIVO

1.1. ESTRATEGIA Y DISEÑO DEL MODELO

Partimos de un *subset* con los tópicos relevantes, el mes que más publicaciones ha tenido cada tópico y el conjunto de noticias publicadas ese mes para dicho tópico, y debemos extraer la información más relevante de esas publicaciones. El primer enfoque que planteamos es:

- 1) **Seleccionamos las primeras frases de cada publicación.** Se hace necesario limitar de alguna manera la longitud de los textos ya que tenemos una limitación del input total a resumir de 512 palabras. Además de que el contenido de las noticias suele sobrepasar con holgura esta longitud, debemos considerar varias publicaciones para el resumen, de hecho todas las publicaciones de ese tópico que se hayan publicado el mes máximo de publicaciones. Decidimos por tanto truncar individualmente el contenido de cada publicación a un número específico que determinaremos. Entendemos que la información más relevante de una noticia se suele explicar al principio, con lo que dicha información estará contenida en el texto extraído.
- 2) **Concatenamos las publicaciones por tópico y mes.** Una vez hemos truncado el contenido de las publicaciones, las concatenamos. Nuestra intuición nos lleva a pensar que el contenido relevante se repetirá y será más evidente.
- 3) **Hacemos un resumen abtractivo del texto concatenado.**

1.2. IMPLEMENTACIÓN

Pasamos a implementar el modelo que hemos diseñado.

Construimos una función que, para un tópico dado, devuelva el año y el mes que más publicaciones ha tenido y un texto que concatene las 4 primeras frases de las publicaciones de ese mes.

```
def generate_relevant_fact(num):
    tm=temas_per_mes[temas_per_mes['tema']==num]
    max=tm['id_documento'].max()
    max_mes=tm[tm['id_documento']==max]
    mes=int(max_mes['mes'].min())
    año=int(max_mes['año'].min())
    temas_bert_t=temas_bert[(temas_bert['tema']==num) & (temas_bert['mes']==mes) & (temas_bert['año']==año)]
    news=temas_bert_t.contenido.values.tolist()
    sentence_topic=[]
    for i in range(0, len(news)):
        a=sent_tokenize(news[i], 'spanish')
        sentence_topic=sentence_topic+a[0:4]
    topic_text="".join(sentence_topic)
    return año, mes, topic_text
```

Generamos ahora el código que utiliza esta función para todos los tópicos relevantes identificados.

```
temas_relevantes_seleccionados['tema']=temas_relevantes_seleccionados.index
temas=temas_relevantes_seleccionados.temas.values.tolist()
```

Generación automática de timelines con Transformers

```
año=list()
mes=list()
topic_text=list()
for i in topicos:
    print(i)
    print(mes)
    a, b, c = generate_relevant_fact(i)
    año.append(a)
    mes.append(b)
    topic_text.append(c)
```

```
#Construyo un dataframe con el año, el mes y la selección de texto para calcular y añadir el resumen
facts=pd.DataFrame(zip(year, mes, topic_text), columns=(['año', 'mes', 'texto']))
facts
```

	año	mes	texto
0	2021	1	Desde que el pasado 27 de diciembre Araceli re...
1	2020	10	El Govern estudia la posibilidad de implantar ...
2	2020	4	José Luis San Martín Izcue, médico de Atención...
3	2020	11	El laboratorio estadounidense Moderna ha confi...
4	2020	9	Este lunes arranca el inicio de curso más inci...
5	2020	4	En plena la crisis del coronavirus, muchos can...
6	2020	3	China no registró ningún nuevo contagio local ...
7	2020	9	La cifra oficial de personas fallecidas a caus...
8	2020	10	La Casa Blanca ha anunciado este viernes que e...
9	2020	3	El Ayuntamiento de Madrid ha acordado habilita...
10	2020	4	La presidenta de la Comunidad de Madrid, Isabe...
11	2021	4	La farmacéutica 'lanacel' ha comunicado acta m...

Hacemos el resumen abstractivo con la función `generate_summary()` del modelo de Transformer seleccionado que ya hemos creado previamente.

```
#Hacemos el resumen
facts['summary']=facts['texto'].apply(lambda x: generate_summary(x))
```

Y mostramos el resultado ordenado por año y mes:

```
facts.sort_values(by=(['año', 'mes']))
```

	año	mes	texto	summary
12	2020	3	El coronavirus ha causado en las últimas 24 ho...	Casi 50. 000 personas han sido diagnosticadas ...
13	2020	3	China no registró ningún nuevo contagio local ...	El número de fallecidos se detectó en todo el ...
18	2020	3	Era menester que ocurriese así con todo lo que...	Anabel Pantoja cancela su enlace en Gran Canar...
19	2020	3	Italia ha vuelto a registrar este viernes resu...	El número de fallecidos supera los 1. 000 fall...
23	2020	3	Felipe VI dirige este miércoles 18 de marzo un...	Don Felipe y Letizia conversan en el Palacio d...
42	2020	3	La Comunitat Valenciana ha confirmado este lun...	El número de fallecidos supera los 750 en el ú...
0	2020	4	Buenas noticias en el Wanda Metropolitano.Con ...	La RFEF confirma el acuerdo para jugar cada 72...
11	2020	4	Este domingo se ha celebrado una nueva Confere...	El presidente del Gobierno, Pedro Sánchez, pid...
14	2020	4	Cataluña ha superado los 9.000 muertos por cor...	El número de fallecidos en las últimas 24 hora...
24	2020	4	El Sindicato Médico de Castilla-La Mancha (CES...	El trabajador, que trabaja en la Unidad de Reh...
27	2020	4	Facebook anunció este miércoles unas ganancias...	El gigante de la publicidad en la red social c...
29	2020	4	La crisis del coronavirus está provocando que l...	La Sociedad Protectora de animales de Mataró l...
36	2020	4	Las autoridades de Reino Unido han confirmado ...	El ministro de Sanidad británico, Boris Johnso...
46	2020	4	Confinados en casa, estamos jugando más.Con el...	La industria de los videojuegos se ha unido pa...
53	2020	4	Un bebé nacido el pasado 12 de marzo en un hos...	El recién nacido se contagió a través del cont...
54	2020	4	Las cuatro fases del plan de desescalada que h...	El plan de desescalada será gradual y asimétri...

Realizamos el mismo proceso para el *subset* que obtuvimos filtrando por la temática “Trump”, obteniendo el siguiente resultado:

```
facts.sort_values(by=['año', 'mes'])
```

	año	mes	texto	summary
2	2020	1	En 2015, Estados Unidos firmó junto a Irán (y ...	El Departamento de Defensa de EE UU confirma l...
8	2020	2	El senador Bernie Sanders ha declarado este ju...	El senador por el estado de New Hampshire obti...
6	2020	4	Estados Unidos se convirtió este martes en el ...	El país norteamericano supera el millón de cas...
0	2020	6	Las protestas no cesan en Estados Unidos, dond...	La Guardia Nacional de EE UU ha decretado el t...
9	2020	7	La estrella televisiva Kim Kardashian pidió es...	El rapero Kanye West, diagnosticado de un tras...
7	2020	8	El presidente de EE.UU., Donald Trump, firmó e...	El presidente de EE UU prohíbe cualquier trans...
4	2020	10	El presidente de Estados Unidos, Donald Trump,...	El presidente de EE UU sufre síntomas leves de...
1	2020	11	La noche electoral en EE UU amenaza con ser la...	El presidente de EE UU, Joe Biden, será muy di...
3	2021	1	Representantes de la Cámara Baja de Estados Un...	Representantes de la Cámara alta de EE UU pres...
5	2021	1	YouTube se ha unido a otras plataformas y rede...	La red social Snapchat suspende la cuenta del ...

1.3. ANÁLISIS DE RESULTADOS

Se incluye en el anexo el resultado completo, acontecimientos o hechos relevantes, que hemos obtenido para las temáticas “pandemia” y “Trump” de las diferentes alternativas.

Con objeto de tener información más global y cualitativa, realizamos el siguiente proceso de valoración: Evaluamos cada uno de los registros obtenidos como resultado, valorando los siguientes parámetros:

- **Corrección sintáctica.** Evaluamos la corrección del resultado desde el punto de vista sintáctico
- **Veracidad.** Evaluamos si el hecho que se describe se ajusta o no a la realidad
- **Centralidad.** Evaluamos cómo de bien o mal el resultado obtenido representa el aspecto más relevante de lo acontecido en ese tópico y mes

Las valoraciones se realizarán con la siguiente escala: Mala/Limitada/Buena

En el caso de la alternativa 1 los resultados de esta evaluación son:

Tabla 3: Resultados alternativa 1

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	70%	54%	36%	70%	70%	40%
Limitada	20%	12%	24%	30%	10%	50%
Mala	10%	34%	40%		20%	10%

Fuente: Elaboración propia

Podemos apreciar que, aunque la corrección sintáctica de los resultados suele ser bastante aceptable, los resultados en cuanto a veracidad y centralidad son malos o muy limitados. Si el parámetro centralidad es importante (determina que de verdad estemos mostrando información relevante), el de veracidad es crítico. No podemos permitir un modelo que proporcione información errónea, y los porcentajes son bastante altos.

2. ALTERNATIVA 2: RESUMEN ABSTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN ABSTRACTIVO

2.1. ESTRATEGIA Y DISEÑO DEL MODELO

Con objeto de intentar salvar las debilidades del modelo diseñado en el apartado anterior, construimos una solución con un enfoque diferente:

- 1) **Resumen abstractivo de cada noticia.** Hemos constatado que la limitación de input total de entrada para las tareas de *summarization* (512 tokens), nos está provocando que tengamos sólo una visión parcial de lo acontecido respecto a un tópico en el mes que estamos analizando. Estamos truncando cada publicación a las 4 primeras frases, y además, aún así, cuando concatenamos tenemos que volver a truncar por exceder la longitud máxima, perdiendo directamente la información de algunas publicaciones. Para intentar solucionar este tema probamos a realizar como paso previo un resumen abstractivo de cada publicación. Con este paso conseguimos limitar la longitud de cada noticia, pero intentando asegurar que conservamos la información clave de cada publicación.
- 2) **Concatenamos las publicaciones por tópico y mes.** Igual que en el modelo diseñado en la alternativa 1, pero en este caso será una concatenación de resúmenes.
- 3) **Hacemos un resumen abstractivo del texto concatenado.**

2.2. IMPLEMENTACIÓN

Modificamos la función que genera hechos relevantes dado un tópico para incorporar el nuevo diseño.

```
def generate_relevant_fact_2(num):
    tm=topicos_per_month[(topicos_per_month['topic']==num)]
    max=tm['id_document'].max()
    max_month=tm[tm['id_document']==max]
    mes=int(max_month['mes'].min())
    año=int(max_month['año'].min())
    topics_bert_t=topics_bert[(topics_bert['topic']==num) & (topics_bert['mes']==mes) & (topics_bert['año']==año)]
    news=topics_bert_t.contenido.values.tolist()
    sentence_topic=[]
    for i in range(0, len(news)):
        sentence_topic=sentence_topic+[generate_summary(news[i])]
    topic_text=" ".join(sentence_topic)
    return año, mes, topic_text
```

Y procedemos, igual que en el punto anterior, a generar el conjunto de hechos relevantes para todos los tópicos que se han identificado como tal.

Obtenemos nuevos resultados:


```
#Hacemos el resumen
facts2['summary']=facts2['texto'].apply(lambda x: generate_summary(x))
facts2.sort_values(by=['año', 'mes'])
```

	año	mes	texto	summary
12	2020	3	El número de casos se eleva a 462. 684, según ...	La OMS cree que la pandemia ha provocado la mu...
13	2020	3	Las autoridades sanitarias confirman que no se...	Casi 1. 400 personas han sido dados de alta tr...
18	2020	3	Anabel Pantoja cancela su boda con Omar Sánche...	Anabel Pantoja y Omar Sánchez posponen en un p...
19	2020	3	El presidente de Lombardía, Silvio Brusafarro,...	El número de muertos con el virus en Italia as...
23	2020	3	Felipe VI dirige este miércoles un mensaje tel...	Don Felipe y doña Letizia conversan con los re...
42	2020	3	La Comunidad Valenciana confirma que la Genera...	La Comunidad Valenciana contabiliza 1. 105 pac...
0	2020	4	Los colchoneros ocupan la sexta plaza de la La...	Una selección de las historias de actualidad d...
11	2020	4	El presidente de la Generalitat, Quim Torra, p...	El presidente del Gobierno, Quim Torra, pide a...
14	2020	4	La cifra de fallecidos en las últimas 24 horas...	El número de fallecidos en las últimas 24 hora...

2.3. ANÁLISIS DE RESULTADOS

Los resultados completos de nuestras temáticas, “pandemia” y “Trump” se han añadido al anexo. Los resultados del proceso de evaluación de estos resultados es el siguiente:

Tabla 4: Resultados alternativa 2

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	67%	67%	28%	70%	40%	30%
Limitada	26%	16%	38%		10%	30%
Mala	7%	17%	34%	30%	50%	40%

Fuente: Elaboración propia

Vemos que no sólo no hemos mejorado, si no que hemos empeorado. El 70% de centralidad es mala y limitada y respecto a la veracidad, en el caso de la temática de “Trump” tiene un 60% de resultados en los que es mala o limitada; es decir, más de la mitad de los hechos identificados aportan información errónea.

3. ALTERNATIVA 3: RESUMEN EXTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN ABSTRACTIVO

3.1. ESTRATEGIA Y DISEÑO DEL MODELO

Abordamos ahora un enfoque bastante diferente. Seguimos intentando reducir la longitud de los inputs a resumir sin perder la información relevante, pero ahora, en lugar de utilizar el resumen abstractivo lo hacemos con técnicas extractivas. Estas técnicas devuelven frases literales del texto, aquellas que identifican como más relevantes. Para ello miden la similitud entre frases (aplicando *cosine similarity*) una vez han representado la información semántica de cada frase en un espacio vectorial multidimensional. Me resulta muy interesante este concepto ya que identificar la similitud entre frases me puede ayudar a identificar qué información relevante es la que se *repite* en todas las publicaciones de un pico temporal.

Diseño la siguiente solución:

- 1) **Realizamos un resumen extractivo de cada noticia.** Para ello, tal y como explicamos en el capítulo IV donde realizamos un estudio del estado del arte, utilizamos Sentence Transformers. Debemos indicar cuántas frases relevantes queremos obtener en cada resumen y especificamos 2.
- 2) **Concatenamos las publicaciones por tópico y mes.** Una vez hemos truncado el contenido de las publicaciones, las concatenamos. En este caso, el hecho de que el contenido relevante se repita y sea más evidente puede ser más probable.
- 3) **Hacemos un resumen extractivo del texto concatenado.**

3.2. IMPLEMENTACIÓN

Realizamos las modificaciones de la función `generate_relevant_facts()` y definimos una función específica para realizar el resumen extractivo que se llamará `ext_summ()`

```
import LexRank
import nltk
from sentence_transformers import SentenceTransformer, util
import numpy as np
from LexRank import degree centrality_scores
```

```
model_sentence_transformer = SentenceTransformer('sentence-transformers/all-MiniLM-L6-v2')
```

```
def ext_summ(text, num):
    sentences = nltk.sent_tokenize(text)
    embeddings = model_sentence_transformer.encode(sentences, convert_to_tensor=True)
    cos_scores = util.cos_sim(embeddings, embeddings).numpy()
    centrality_scores = degree_centrality_scores(cos_scores, threshold=None)
    most_central_sentence_indices = np.argsort(-centrality_scores)
    ext_summary=""
    for i in range(0, num):
        ind=most_central_sentence_indices[i]
        ext_summary=ext_summary+" "+sentences[ind]
    return ext_summary
```

```
def generate_relevant_fact(num):
    tm=topicos_per_month[(topicos_per_month['topic']==num)]
    max=tm['id_document'].max()
    max_month=tm[tm['id_document']==max]
    mes=int(max_month['mes'].min())
    año=int(max_month['año'].min())
    topics_bert_t=topics_bert[(topics_bert['topic']==num) & (topics_bert['mes']==mes) & (topics_bert['año']==año)]
    news=topics_bert_t.contenido.values.tolist()
    sentence_topic=[]
    for i in range(0, len(news)):
        a=ext_summ(news[i], 1)
        sentence_topic.append(a)
    topic_text="".join(sentence_topic)
    return año, mes, topic_text
```

```
topicos_relevantes_selected['topico']=topicos_relevantes_selected.index
topicos=topicos_relevantes_selected.topico.values.tolist()
```

```
año=list()
mes=list()
topic_text=list()
for i in topicos:
    a, b, c = generate_relevant_fact(i)
    año.append(a)
    mes.append(b)
    topic_text.append(c)
```

```
#Construyo un dataframe con el año, el mes, la selección de texto y el resumen
facts=pd.DataFrame(zip(año, mes, topic_text), columns=(['año', 'mes', 'texto']))
facts
```

```
#Hacemos el resumen
facts['summary']=facts['texto'].apply(lambda x: generate_summary(x))
```

```
facts.sort_values(by=(['año', 'mes']))
```

	año	mes	texto	summary
12	2020	3	La propagación más rápida del virus está ocur...	El número de casos de esta enfermedad supera l...
13	2020	3	La Comisión Nacional de Sanidad señaló que, h...	El número de fallecidos en la enfermedad se re...

3.3. ANÁLISIS DE RESULTADOS

Como en el resto de los apartados, los resultados completos de las temáticas “pandemia” y “Trump” se han incluido en el anexo.

Analizamos en este apartado la evaluación cualitativa realizada sobre los resultados.

Tabla 5: Resultados alternativa 3

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	66%	50%	17%	60%	60%	20%
Limitada	10%	10%	41%	10%	10%	30%
Mala	24%	40%	41%	30%	30%	50%

Fuente: Tabla de elaboración propia

Los resultados no arrojan tampoco una mejora sustancial. De hecho, empeoran casi todos los parámetros evaluados. Definitivamente, combinar resumen extractivo con abstractivo no nos ha funcionado. En el primer paso, cuando hacemos el resumen extractivo, obtenemos dos noticias por publicación que muchas veces aportan información inconexa entre ellas. Esto motiva que el modelo de resumen abstractivo pierda todo el contexto y no aporte buenos resultados.

4. ALTERNATIVA 4: RESUMEN EXTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN EXTRACTIVO

4.1. ESTRATEGIA Y DISEÑO DEL MODELO

Probamos por último un nuevo diseño que nos ayude a salvar algunas carencias identificadas hasta el momento. En este caso, realizamos el resumen que abordamos como último paso extractivo en lugar de abstractivo.

4.2. IMPLEMENTACIÓN

Reutilizo las funciones e instrucciones de la alternativa 3, debo cambiar únicamente el último resumen. De hecho, lo añado como una columna nueva al resultado de dicha alternativa. Indico a la función que me genera el resumen extractivo que me devuelva las dos frases más relevantes de la concatenación de frases relevantes de las publicaciones.

```
#Hacemos el resumen extractivo
facts['summary2']=facts['texto'].apply(lambda x: ext_summ(x, 2))

facts.sort_values(by=(['año', 'mes']))
```

	año	mes	texto	summary	summary2
12	2020	3	La propagación más rápida del virus está ocur...	El número de casos de esta enfermedad supera l...	En relación a los tratamientos, Tedros ha des...
13	2020	3	La Comisión Nacional de Sanidad señaló que, h...	El número de fallecidos en la enfermedad se re...	Italia, sin embargo, empieza este jueves una ...
18	2020	3	En este grupo se encuentra precisamente Anabe...	Anabel Pantoja, de 71 años, canceló su boda co...	En este grupo se encuentra precisamente Anab...
19	2020	3	Por otra parte, el presidente del Instituto S...	El presidente del Instituto Superior de Salud ...	Por otra parte, el presidente del Instituto ...
23	2020	3	Felipe VI dirige este miércoles 18 de marzo u...	El presidente del Cermi asegura que las person...	Los Reyes también han mantenido una videoconf...
42	2020	3	Además, 73 mayores de estas instalaciones han...	La Consejería de Sanidad ha confirmado la muer...	El centro se ha reforzado con 20 enfermeros, ...
0	2020	4	Con la incertidumbre del parón futbolístico y...	Los jugadores de LaLiga no pasan el confinamie...	Con la incertidumbre del parón futbolístico ...
11	2020	4	La presidenta de la Comunidad de Madrid, Isab...	La presidenta de la Comunidad de Madrid nide a	El presidente del Gobierno Pedro Sánchez ha a

4.3. ANÁLISIS DE RESULTADOS

Los resultados completos de la alternativa 4 para las temáticas “pandemia” y “Trump” se incluyen en el anexo.

La evaluación de esta alternativa arroja los siguientes resultados:

Tabla 6: Resultados alternativa 4

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	84%	88%	41%	60%	100%	80%
Limitada	14%	12%	45%	40%		20%
Mala	2%		14%			

Fuente: Elaboración propia

Comprobamos que mejoramos notablemente en todos los parámetros medidos a expensas de, si bien es cierto, descripciones demasiado extensas y no siempre perfectamente coherentes. Lo que hacemos en este método es extraer frases relevantes de una lista o concatenación de las dos frases más relevantes de cada publicación. En la mayoría de los casos ambas se complementan, aportan información más precisa, pero no siempre tienen mucho sentido sintáctico y/o semántico estando juntas.

Un parámetro clave en este diseño es el número de frases relevantes que nos devuelve el proceso de *summarization*. El modelo funciona mucho mejor cuando extraemos dos frases, tanto en el paso previo de resumen de cada publicación, como en el resumen final por tópico/mes. El motivo es que una única frase en algunas ocasiones no representa suficientemente bien el hecho central de ese tópico/mes. En estos casos, una segunda frase complementa la información o, incluso la aporta si la primera era irrelevante. Hay casos muy concretos en los que la primera frase no es relevante, pero la segunda sí. Como se puede comprobar en los resultados del parámetro de evaluación "centralidad", con dos frases logramos resultados muy buenos.

Es cierto, como ya he comentado, que conseguimos valores realmente aceptables en las métricas de "veracidad" y "centralidad" a costa de resúmenes demasiado largos y no siempre perfectamente coherentes. Mostramos a continuación, un extracto del resultado (incluido de manera completa en el anexo) para que se puedan valorar los aspectos que acabo de exponer.

Temática "pandemia":

Año: 2020 Marzo

- Además, Tedros ha subrayado la importancia de hacer frente al nuevo coronavirus con iniciativas "agresivas y específicas", entre las que ha destacado la realización de la prueba a toda persona sospechosa de padecerlo, asilar y tratar a los casos confirmados y poner en cuarentena a las personas que hayan tenido un contacto estrecho con un infectado. La propagación más rápida del virus está ocurriendo sin duda en Estados Unidos, que está cerca de los 65.000 casos, lo que impli

ca más de 11.000 en un solo día, mientras que las muertes ya han superado el millar.

- En España, el número de contagios ha llegado este miércoles a superar los 2.100 casos y 55 personas han muerto con coronavirus. Así, hasta el viernes, se han registrado un total de 81.008 casos confirmados de coronavirus en la China continental y 3.255 personas han muerto por la enfermedad en el país asiático.

RESULTADOS Y DISCUSIÓN / DESARROLLO DE LA ARGUMENTACIÓN

Si atendemos a los resultados obtenidos en las distintas alternativas o soluciones implementadas durante este trabajo de fin de máster observamos que:

- Hemos sido capaces de construir un modelo que bucea en los contenidos de 95.008 publicaciones de un periódico digital, extrayendo un resumen datado de acontecimientos relativos a una temática concreta
- En las temáticas que hemos analizado en profundidad, “pandemia” y “Trump”, hemos obtenido resúmenes de aspectos clave acontecidos, con más o menos acierto según la alternativa implementada. En el caso de la temática “pandemia” extrayendo información relevante de 17.205 publicaciones y 1.920 en el caso de “Trump”
- En la última solución implementada, que es la que mejores resultados ha arrojado, vemos reflejados gran parte de los acontecimientos clave:
 - “pandemia”: expansión de la pandemia en marzo desde China, rastreos, paralización/activación de comercios minoristas, infravaloración inicial (marzo) de lo que ha supuesto la pandemia, parón del fútbol y la Liga, desescalada, ERTes, fondos de recuperación europeos...
 - “Trump”: impacto de la pandemia en EEUU, protestas por los derechos afroamericanos, prohibición de TikTok (popular red social china) en EEUU, positivo de Trump, elecciones en EEUU, *impeachment*, cancelación de las redes sociales de las cuentas de Trump...
- Nos encontramos por otro lado algunos acontecimientos identificados como clave que realmente no lo son, como por ejemplo la cancelación de la boda de Anabel Pantoja en marzo 2020
- Echamos en falta algunos hechos relevantes que no están claramente identificados, como puede ser el confinamiento de toda la población española en marzo del 2020, la declaración del estado de alarma...
- Encontramos algunos hechos relevantes que son demasiado específicos, como aquellos que dan cifras concretas en datos que se han estado actualizando casi a diario, o los referidos a una única comunidad autónoma
- Y, refiriéndonos al aspecto más formal, obtenemos unos resultados o resúmenes demasiado largos, poco concretos, sin capacidad para generalizar

Analizamos cada uno de los procesos de nuestra solución para valorar de manera global las diferentes alternativas que hemos implementado e identificar mejoras en el diseño del modelo. La solución consta de los siguientes procesos:

- Extracción de publicaciones relativas a la temática
- Modelado de tópicos
- Identificación de tópicos relevantes
- Extracción de información relevante

1) Extracción de publicaciones relativas a la temática

En este punto hemos simplificado el proceso de extracción limitándonos a hacer un filtro en los contenidos de las publicaciones con la o las palabras clave a analizar. Este será seguramente uno de los principales motivos por los que no hemos identificado algunos acontecimientos importantes. Filtrar por “pandemia” excluye aquellas noticias que no referencia exclusivamente a este término pero sí a otros muy relacionados con esta temática como son “covid” o “coronavirus”.

Una mejora a implementar sería realizar un proceso de filtrado más sofisticado que incluyera los términos muy similares por contexto al introducido como temática. Esto se puede hacer realizando un análisis previo de todo el *dataset* que calcule estas similitudes.

2) Modelado de tópicos

El modelado de tópicos ha funcionado bastante bien, a pesar de haber identificado algunas temáticas muy concretas en las que se han identificado un número de tópicos muy bajo.

El aspecto más importante que considerar en esta parte del proceso es el tiempo de ejecución para el entrenamiento del modelo y la obtención de los vectores de tópicos y palabras. Según el análisis que hemos realizado con diferentes temáticas (tanto en número total de publicaciones como en ámbito o contexto de la temática), tenemos dos opciones:

- Utilizar un modelo pre-entrenado para el modelado de tópicos
- Utilizar doc2vec, que entrena un modelo desde cero

Si pensamos en una solución que se pueda implementar como funcionalidad web o de una app (periódico digital), los tiempos de la segunda opción no son asumibles cuando el *subset* de publicaciones a modelar es voluminoso, y los resultados del modelado de tópicos con los modelos pre-entrenados es suficientemente buena. Por otro lado, en *subsets* más pequeños los tiempos sí pueden ser asumibles y vemos una mejora para algunos casos en los que la identificación de tópicos es pobre. Estos son los resultados que obtuvimos:

Tabla 7: Resultados Top2Vec con BERT Sentence Transformer

TEMÁTICA	TOTAL PUBLICACIONES	Nº TÓPICOS	TIEMPO DE EJECUCIÓN (min.)
Pandemia	17.205	91	20
Trump	1.920	15	4
China	3.167	2	6
5G	542	7	2
Elecciones	3.955	3	8

Fuente: Elaboración propia

Tabla 8: Resultados Top2Vec con doc2vec

TEMÁTICA	TOTAL PUBLICACIONES	Nº TÓPICOS	TIEMPO DE EJECUCIÓN (min.)
Pandemia	17.205	180	222
Trump	1.920	21	8
China	3.167	2	21
5G	542	5	2
Elecciones	3.955	53	27

Fuente: Elaboración propia

Optaríamos en nuestra solución por definir en un parámetro un número máximo de publicaciones por temática, eligiendo la opción BERT Sentence Transformer para temáticas que superen este número y doc2vec para el resto.

3) Identificación de tópicos relevantes

Para obtener los tópicos relevantes he adoptados el siguiente criterio: tópicos que han tenido un número de publicaciones mensuales superior a la media de publicaciones tópico/mes.

Este criterio considera únicamente el volumen de publicaciones y al tener un único medio digital para analizar, lleva a situaciones como la de destacar como hecho relevante algunos acontecimientos más relacionados por ejemplo con el ámbito social (boda de Anabel Pantoja). En este sentido, enriquecería mucho nuestra solución contar con un *dataset* que integrara las publicaciones de varios medios digitales, ya que permitiría comparar y proporcionar una visión más objetiva.

4) Extracción de información relevante

Hemos implementado y analizado cuatro soluciones diferentes para la extracción de información relevante, conjugando opciones de resumen abstractivo y extractivo. Todas ellas con modelos Transformers.

El enfoque de solución abstractivo y el extractivo son radicalmente diferentes. El primero procesa la información, obteniendo contexto y entendimiento del contenido, y genera un texto con lo que ha identificado como relevante. Este texto no coincidirá literalmente con el contenido origen. El segundo en cambio representa vectorialmente las frases e identifica cuáles son más relevantes aplicando métricas de similitud (*cosine similarity*). El resultado serán frases literales del contenido origen.

Para poder evaluar los diferentes métodos hemos analizado dos temáticas muy diversas tanto en volumen de publicaciones como en ámbito y contexto, y hemos clasificado los resultados, valorando los siguientes parámetros en una escala cualitativa: Buena/Limitada/Mala

- **Corrección sintáctica.** Evaluamos la corrección del resultado desde el punto de vista sintáctico
- **Veracidad.** Evaluamos si el hecho que se describe se ajusta o no a la realidad
- **Centralidad.** Evaluamos cómo de bien o mal el resultado obtenido representa el aspecto más relevante de lo acontecido en ese tópico y mes

Los resultados han sido:

Tabla 9: Resultados alternativa 1 - CONCATENACIÓN DE NOTICIAS Y RESUMEN ABSTRACTIVO

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	70%	54%	36%	70%	70%	40%
Limitada	20%	12%	24%	30%	10%	50%
Mala	10%	34%	40%		20%	10%

Fuente: Elaboración propia

Tabla 10: Resultado alternativa 2 - RESUMEN ABSTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN ABSTRACTIVO

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	67%	67%	28%	70%	40%	30%
Limitada	26%	16%	38%		10%	30%
Mala	7%	17%	34%	30%	50%	40%

Fuente: Elaboración propia

Tabla 11: Resultados alternativa 3 – RESUMEN EXTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN ABSTRACTIVO

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	66%	50%	17%	60%	60%	20%
Limitada	10%	10%	41%	10%	10%	30%
Mala	24%	40%	41%	30%	30%	50%

Fuente: Elaboración propia

Tabla 12: Resultados alternativa 4 - RESUMEN EXTRACTIVO POR NOTICIA, CONCATENACIÓN Y RESUMEN EXTRACTIVO

	Temática: "pandemia"			Temática: "Trump"		
	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD	CORR. SINTÁCTICA	VERACIDAD	CENTRALIDAD
Buena	84%	88%	41%	60%	100%	80%
Limitada	14%	12%	45%	40%		20%
Mala	2%		14%			

Fuente: Elaboración propia

A la vista de los resultados sacamos algunas conclusiones:

- Analizando la alternativa 1, que es la que implementa el modelo Transformer para tareas abstractivas con menos manipulación previa, los parámetros de veracidad y centralidad no son todo lo buenos que requerimos. Soy especialmente estricta con el de veracidad, ya que no podemos diseñar una funcionalidad que aporte información que no es real. Es importante remarcar que en esta evaluación no estamos valorando el modelo Transformer en sí, ya que, gran parte de los problemas que nos encontramos se deben a que concatenamos información de varias publicaciones y, como tenemos el input para la tarea de resumen limitado a 512 tokens, hemos implementado estrategias de truncado que aportan sólo información parcial, truncado en cada publicación (sólo las 4 primeras frases), y truncado del concatenado final de publicaciones, que en muchas ocasiones supera este límite y nos obliga a dejar publicaciones fuera del análisis
- Con objeto de mejorar los parámetros de veracidad y centralidad de la alternativa 1 hemos ido abordando diversas estrategias que permitan reducir el input conservando mejor la información relevante (y mejorando por tanto los parámetros de veracidad y centralidad finales) sin éxito. Las métricas no sólo no han mejorado, si no que han empeorado. Concluimos que el resumen abstractivo no funciona bien con inputs que concatenan diversos resúmenes previos, sean extractivos o abstractivos. La coherencia del input parece ser importante
- Finalmente, hemos implementado una estrategia de resumen extractivo sobre la concatenación de resúmenes previos extractivos que ha mejorado sustancialmente todos los parámetros. Tiene como gran aspecto a mejorar que el resumen final, nuestros hechos relevantes, son extensos y no siempre explicados con completa coherencia (son dos frases literales de las publicaciones origen identificadas como las más relevantes)

Como aspectos de mejora en el diseño de nuestra solución de este último proceso destacaría:

- Explorar otros modelos de Transformers para el resumen abstractivo que se basaran en un modelo pre-entrenado que permitiera un input de más longitud. Los hay que permiten hasta 1.024 tokens. Yo he utilizado un modelo pre-entrenado específicamente para tareas de resumen abstractivo en español. No he encontrado ninguno pre-entrenado con un modelo de 1.024 tokens como input, así que tendría que realizar *fine-tuning* y entrenar mi propio modelo de *summarization*!
- Otra mejora específica que necesita mi solución es conseguir que los resúmenes sean capaces de generalizar lo suficiente (problemas que hemos tenido con cifras concretas, información de una ciudad o comunidad autónoma concreta...). Volvemos a necesitar realizar *fine-tuning* y entrenar específicamente un modelo Transformer para tareas de *summarization*

CONCLUSIONES FINALES

En este trabajo fin de máster he abordado un reto complejo que me ha llevado a explorar varios caminos, muchos de ellos frustrantes en los resultados que iba obteniendo, aunque otros bastante sugerentes y motivadores. Este trabajo de más de dos meses explorando técnicas y herramientas del ámbito NLP me permite obtener algunas conclusiones:

- El modelado de tópicos es una herramienta imprescindible para analizar temas, conversaciones... activando iniciativas de *social listening*. Nos permite dar ese gran paso de gigante que supone llegar a entender qué temas preocupan y ocupan a las personas en este enorme universo de conversaciones digitales
- Los modelos de resumen abstractivo construidos con Transformer tienen una corrección sintáctica que, siendo modelos generativos del lenguaje, es muy sorprendente
- Los modelos de resumen extractivo funcionan muy bien. Son una herramienta muy poderosa para la extracción de información relevante de cantidades muy voluminosas de texto
- En ambos casos, modelos Transformers, la limitación del *input* de entrada o texto a resumir es un hándicap. En los modelos que hemos utilizado para las soluciones que hemos implementado en este TFM 512 *tokens*. Hay modelos que admiten hasta 1024, pero seguimos estando limitados. Esto nos lleva a tener que diseñar estrategias para gestionar grandes volúmenes de texto con más o menos éxito, y que siempre desvirtúan y empeoran el funcionamiento general de Transformers con textos más cortos
- A pesar de las limitaciones que hayamos podido encontrar, los modelos Transformers están sin duda revolucionando las capacidades en el ámbito NLP y van a suponer un nuevo paradigma. Los modelos previos (Redes Neuronales) eran increíblemente costosos de entrenar, y el hardware ha sido siempre una limitación complicada. Por dos motivos fundamentales: primero, que pocos tienen realmente la capacidad de hardware necesaria para abordar tareas complejas, con lo que limitamos mucho la contribución e investigación, y en segundo lugar, que el costo en términos económicos y medioambientales que supone el gasto computacional de los modelados previos a Transformers era poco asumible a medio/largo plazo. El concepto introducido por estos nuevos modelos parte de reutilizar, concepto imprescindible en los tiempos que estamos viviendo actualmente. Se entrenan con sistemas hardware potentísimos modelos que después podemos reutilizar todos los que queramos construir funcionalidades NLP con un coste computacional bajísimo
- Terminar con una reflexión. Una de las motivaciones cuando abordé este TFM era centrarme en el diseño de alguna solución en el ámbito de NLP en español, ya que quería estudiar en profundidad el estado del arte en español. Tal y como sospechaba, tanto en lo que respecta a la disponibilidad de *datasets* para entrenar modelos, como disponibilidad de modelos pre-entrenados tenemos grandes carencias. De hecho, tuve que construir mi

propio *dataset* para este trabajo y el número de modelos de *summarization* en idioma español que he encontrado ya entrenado son muy pocos (aunque muy buenos!). Desde que surgieron los Transformers tenemos el camino mucho más llano para construir y contribuir al ecosistema de la IA en español. Es un camino que debemos recorrer sin ninguna demora, no nos podemos quedar atrás en esta carrera, y nunca había sido tan accesible y asequible diseñar soluciones del ámbito NLP. Soñar, visualizar y diseñar. El único límite no es todavía la imaginación, pero casi...

BIBLIOGRAFÍA

- [1] Rob van Zoes. (10 de Diciembre de 2020). *Start your NLP journey with this Periodic Table of 80+ NLP tasks*. Medium. <https://medium.com/innerdoc/80-natural-language-processing-tasks-described-c777bc4974b3>
- [2] Aravind Pai. (2020). Text Summarization using Deep Learning. *International Journal of Recent Technology and Engineering*, 9(1), 2663–2667.
<https://doi.org/10.35940/ijrte.a3056.059120>
- [3] Kaggle. (). *Datasets*. <https://www.kaggle.com/datasets>
- [4] Hugging Face. (). *Datasets*. <https://huggingface.co/datasets>
- [5] Roesslein, J. (2009). *Tweepy Documentation — tweepy 3.10.0 documentation*. Tweepy Documentation. <https://docs.tweepy.org/en/v3.10.0/index.html>
- [6] T. (2020, 26 diciembre). *GitHub - tweepy/tweepy: Twitter for Python!* GitHub - Tweepy: Twitter for Python! <https://github.com/tweepy/tweepy>
- [7] Zacharias, C. (2018). *GitHub - twintproject/twint: An advanced Twitter scraping & OSINT tool written in Python that doesn't use Twitter's API, allowing you to scrape a user's followers, following, Tweets and more while evading most API limitations*. GitHub. <https://github.com/twintproject/twint>
- [8] Sagar Pundir. (2020). TOP2VEC: New way of topic modelling. *Towards data science*.
Published. <https://towardsdatascience.com/top2vec-new-way-of-topic-modelling-bea165eeac4a>
- [9] Joyce Xu. (2018). Topic Modeling with LSA, PLSA, LDA & lda2Vec. *Medium*.
Published. <https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05>
- [10] Dima Angelov. (2020, agosto). *TOP2VEC: DISTRIBUTED REPRESENTATIONS OF TOPICS*. <https://arxiv.org/pdf/2008.09470.pdf>

[11] Hugging Face. (s. f.). *Hugging Face – The AI community building the future.*

<https://huggingface.co/>. Recuperado 21 de septiembre de 2021, de

<https://huggingface.co/>

[12] *Transformer models - Hugging Face Course.* (s. f.). <https://huggingface.co/>.

Recuperado 21 de septiembre de 2021, de <https://huggingface.co/course/chapter1>

[13] Reimers, Nils And Gurevych, & Iryna. (2020). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.* Association for

Computational Linguistics. <https://arxiv.org/abs/2004.09813>

[14] Manuel Romero. (s. f.). *Spanish BERT2BERT (BETO) fine-tuned on MLSUM ES for summarization.* <https://huggingface.co/>. Recuperado 21 de septiembre de 2021, de

https://huggingface.co/mrm8488/bert2bert_shared-spanish-finetuned-summarization

ANEXOS

Mostramos a continuación los resultados completos de las diferentes alternativas utilizadas para la temática “pandemia” y “Trump”.

1. RESULTADOS ALTERNATIVA 1

Temática “pandemia”

Año: 2020 Marzo

- Casi 50. 000 personas han sido diagnosticadas en todo el mundo en las 24 horas del día
- El número de fallecidos se detectó en todo el país en la víspera de l inicio del brote
- Anabel Pantoja cancela su enlace en Gran Canaria, pero no puede celebrarse
- El número de fallecidos supera los 1. 000 fallecidos en 24 horas, según Protección Civil
- Don Felipe y Letizia conversan en el Palacio de la Zarzuela sobre la crisis del coronavirus
- El número de fallecidos supera los 750 en el último balance del domingo

Año: 2020 Abril

- La RFEF confirma el acuerdo para jugar cada 72 horas en la Liga de Campeones
- El presidente del Gobierno, Pedro Sánchez, pide a los presidentes de las comunidades autónomas que faciliten una lista de infraestructuras
- El número de fallecidos en las últimas 24 horas supera los 3. 756 personas
- El trabajador, que trabaja en la Unidad de Rehabilitación de Área, falleció en el Hospital Virgen de la Salud
- El gigante de la publicidad en la red social confirma que el aumento de las ventas de la empresa de publicidad es menos mala de la esperada, e incluso permite vislumbrar un cierto regreso a la estabilidad

- La Sociedad Protectora de animales de Mataró lanza una campaña para llamar a los ciudadanos de las protectoras de animales
- El ministro de Sanidad británico, Boris Johnson, confirma que se ha superado el pico de la pandemia en el país
- La industria de los videojuegos se ha unido para promover los mensajes de la Organización Mundial de la Salud contra el coronavirus
- El recién nacido se contagió a través del contacto estrecho con su madre y no a través de la placenta
- El plan de desescalada será gradual y asimétrica, según el Gobierno

Año: 2020 Mayo

- La cifra de fallecidos en las 24 horas supera los 3. 000
- El vicepresidente del Gobierno asegura que las nacionalizaciones " son perfectamente posibles en la Constitución española "
- El estado de Nueva York se mantiene como el epicentro de la pandemia en EE UU
- El ministro del Interior, José Luis Martínez - Almeida, asegura que es " gravísimo " que el ministro " fulmine " al jefe de la Guardia Civil
- El presidente de Venezuela, Nicolás Maduro, acusa a Colombia y Estados Unidos de actuar en territorio colombiano

Año: 2020 Julio

- Una selección de las historias de actualidad de la jornada
- El número de contagiados supera las 76. 000 personas en las 24 horas de la primera muerte
- Estados Unidos acumula 3, 71 millones de casos tras sumar en las 24 horas 66. 300 casos
- La nueva entrega del consultorio de Psicología @ 20minutos ayuda a resolver las dudas y dificultades que tienen los lectores

Año: 2020 Septiembre

- El curso escolar 2020 - 2021 arranca este lunes en el País Vasco, Valencia, Navarra y La Rioja

- El Gobierno y los sindicatos negocian la prórroga de los ERTE derivados de la crisis

Año: 2020 Octubre

- El director para Europa de la OMS, Hans Kluge, advierte de que la situación de la pandemia en Europa es muy preocupante
- El número de fallecidos supera los 20. 000, según los datos de Salud
- La cifra de fallecidos total se sitúa en 13. 858, 39 más que en el recuento del lunes
- El rey Felipe VI, acompañado de la reina Letizia y sus hijas, recibirá en Oviedo los Premios Princesa de Asturias 2020

Año: 2020 Noviembre

- El presidente de la Junta de Andalucía confirma el cierre de toda actividad no esencial a finales de noviembre
- El ministro de Sanidad, Salvador Illa, se reunirá con las autoridades asturianas para valorar la situación
- Joe Biden, candidato demócrata a la presidencia de EE UU, es el favorito para elegir a su próximo presidente
- En España, más del 70 % de las personas padecen ansiedad, según un estudio de Affor Prevención psicológica
- La prueba, que detecta la presencia del virus en el organismo, permite diagnosticar el nuevo tipo de test

Año: 2020 Diciembre

- La Policía Local de Sevilla desaloja a una veintena de personas que se encontraba en la plaza del Teucro
- El ministro de Sanidad, Ignacio Aguado, advierte de que el inicio de la inmunización aún tardará unos días
- La vacuna contra el coronavirus llegará a España en menos de una semana. Solo tres de cada diez españoles renuncian a vacunarse
- Las medidas de seguridad sanitarias han impedido que este año el sorteo de la Lotería de Navidad haya contado con público
- El Consejo de Ministros aprobará la prohibición de cortar los suministros de agua, agua y gas a quienes no puedan pagarlos

- El Instituto de Astrofísica de Canarias (IAC), en Tenerife (Tenerife), ha informado este miércoles del fenómeno astronómico

Año: 2021 Enero

- La pandemia de coronavirus, la más taquillera en España, vuelve a retrasar su estreno en 2020

Año: 2021 Febrero

- La inmunización de la vacuna contra la enfermedad de coronavirus es el paso definitivo de la inmunización mundial

- El presidente del PSC, Pere (ERC), ha sido el protagonista de la polémica de las últimas jornadas de campaña electoral

- La pandemia de Covid - 19 ha provocado que aumenten los casos de jóvenes que ejercen violencia sobre sus padres y, en algunas ocasiones, se agraven los trastornos mentales

- Científicos de la Universidad de Guadalajara investigan si las mutaciones halladas en cuatro casos de coronavirus pueden condicionar la efectividad de las vacunas

Año: 2021 Marzo

- La ministra de Asuntos Exteriores, Arancha González Laya, asegura que el Gobierno " siempre lo va a hacer "

- El 8 de marzo, Día Internacional de la Mujer, estará marcado por la crisis sanitaria de la Covid - 19 y no habrá manifestaciones masivas

- Una encuesta del CIS revela que el sistema de atención a las empresas en 2021 es positivo para las empresas

- La velocidad de reproducción del virus aumenta tres centésimas más que el pasado mes de febrero

- El plazo para tramitar la declaración de la Renta de 2020 por Internet comenzó el miércoles 7 de abril de 2021

Año: 2021 Abril

- La Semana Santa de este año vuelve a estar marcada por las restricciones y restricciones de aforo y medidas de seguridad
- La reina británica despidе en Windsor a su marido, fallecido el pasado día 9 de febrero

Año: 2021 Mayo

- El Gobierno de Rajoy eleva el tono ante la crisis migratoria
- Las asociaciones de jueces creen que el Gobierno no ha legislado adecuadamente en los últimos nueve meses
- La nit dels Museus cuenta con más de 2. 600 participantes para conocer los hábitos de visita y experiencia en los museos
- El país asiático supera los 400. 000 casos en la última jornada de la pandemia

Temática "Trump":

Año: 2020 Enero

- El Departamento de Defensa de EE UU confirma la retirada del acuerdo y la imposición de nuevas sanciones

Año: 2020 Febrero

- El senador por el estado de New Hampshire obtiene el 26, 2 % de los apoyos frente al 20 % del apoyo

Año: 2020 Abril

- El país norteamericano supera el millón de casos confirmados de COVID - 19

Año: 2020 Junio

- La Guardia Nacional de EE UU ha decretado el toque de queda en el centro de la capital estadounidense

Año: 2020 Julio

- El rapero Kanye West, diagnosticado de un trastorno bipolar, se ha visto obligado a ser juzgado en EE UU

Año: 2020 Agosto

- El presidente de EE UU prohíbe cualquier transacción con el desarrollador chino de TikTok

Año: 2020 Octubre

- El presidente de EE UU sufre síntomas leves de la Covid - 19, según ha informado su médico

Año: 2020 Noviembre

- El presidente de EE UU, Joe Biden, será muy difícil adelantar esta noche el escrutinio en otro Estado clave

Año: 2021 Enero

- Representantes de la Cámara alta de EE UU presentan la acusación contra el expresidente
- La red social Snapchat suspende la cuenta del presidente de EE UU tras ser suspendida por la compañía

2. RESULTADOS ALTERNATIVA 2

Temática "pandemia":

Año: 2020 Marzo

- La OMS cree que la pandemia ha provocado la muerte de más de 50 millones de personas en Europa. El número de casos se eleva a 462. 684

- Casi 1. 400 personas han sido dados de alta tras ser diagnosticadas en Hong Kong, la provincia más afectada
- Anabel Pantoja y Omar Sánchez posponen en un programa de Telecinco para hablar de la cancelación del enlace
- El número de muertos con el virus en Italia asciende a 74. 386, según Protección Civil
- Don Felipe y doña Letizia conversan con los representantes del Comité Español de Personas con Discapacidad (Cermidad)
- La Comunidad Valenciana contabiliza 1. 105 pacientes, de los cuales 378 se encuentran en unidades de cuidados intensivos

Año: 2020 Abril

- Una selección de las historias de actualidad de la jornada
- El presidente del Gobierno, Quim Torra, pide a los presidentes de las comunidades autónomas que faciliten una lista de infraestructuras
- El número de fallecidos en las últimas 24 horas supera los 9. 923 en las 24 horas
- El trabajador, que había contraído coronavirus hace un año, había lanzado varios mensajes de esperanza en la sociedad
- La red social elimina grupos y eventos que alienten a las personas a desafiar la orientación del distanciamiento social
- Big Poppa, un bulldog que vive en Atlanta junto a Rae Elle, es uno de los muchos ejemplos de cómo la cuarentena afecta a las mascotas
- El ministro británico de Exteriores, Rishi Sunak, alerta del peligro de una segunda ola sin control de contagios
- La venta de la consola de Nintendo Switch se ha convertido en objeto de deseo por la compañía
- El recién nacido, de cuatro días, ha sido diagnosticado con insuficiencia respiratoria
- El plan de desescalada se realizará en mayo cuando el país entre en la fase 1, que estará prevista en mayo

Año: 2020 Mayo

- En los últimos siete días, 995 personas han precisado de hospitalización en la última semana

- El vicepresidente del Banco Central Europeo cree que España, Portugal e Italia no entran en detalles de la renta mínima
- El estado de Nueva York se mantiene como epicentro de la pandemia en Estados Unidos
- El ministro del Interior destituye al jefe de la Comandancia de la Benemérita
- El presidente de Venezuela, que compartía de facto desde enero con Parra, confirma que el presidente de la Asamblea de Justicia anuló la presidencia del organismo

Año: 2020 Julio

- Una selección de las historias de actualidad de la jornada
- El número de fallecidos supera los 1. 022. 055, según el Gobierno
- El país asiático supera los 4, 42 millones de casos con 150. 000 personas en las 24 horas
- La psicóloga Magda Barceló explica las claves para lograr el empleo estable en las agencias de colocación

Año: 2020 Septiembre

- El consejero de Educación, Jokin Bildarratz, analiza el desarrollo de las reuniones con agentes de la comunidad
- La ministra de Trabajo, Yolanda Díaz, se muestra partidaria de alargar la vigencia de los ERTE

Año: 2020 Octubre

- La especialista en enfermedades emergentes alerta de la expansión del virus en toda Europa y pide a los Estados miembros a "intensificar" su respuesta
- El número de fallecidos supera los 33. 037, con 104 muertes confirmados en el último día
- La cifra de fallecidos total se sitúa en 13. 819, 38 más que en el registro anterior
- Doña Leonor alaba la actitud de los jóvenes ante la pandemia de coronavirus

Año: 2020 Noviembre

- El Boletín Oficial de la Junta de Andalucía publica las nuevas restricciones por la pandemia del coronavirus
- El ministro de Sanidad, Salvador Illa, asegura que con las medidas incluidas en el nuevo estado de alarma se podrá controlar la evolución de la pandemia
- El presidente de EE UU, Joe Biden, se declara " fraude " en las elecciones de enero
- El 70 % de las personas de entre 18 y 39 años son encuestadas en un estudio de 20 minutos
- El test de EE UU para autodiagnóstico en casa permite detectar el nuevo coronavirus SARS

Año: 2020 Diciembre

- El cuerpo policial intervino, a petición de un ciudadano, en un bar de la barriada de Triana
- El Consejo Interterritorial de Sanidad analizará los resultados clínicos de la vacuna de Pfizer
- Solo tres de cada diez españoles renuncian a vacunarse en cuanto puedan, tal como se extrae del barómetro del CIS de diciembre
- El sorteo de la Lotería de Navidad ha marcado el inicio de la fiesta navideña
- El Gobierno estudia cómo compensar los desahucios mientras dure el estado de alarma por el desahucio
- El Instituto de Astrofísica de Canarias (IAC) y el Instituto Geográfico Nacional emiten el fenómeno

Año: 2021 Enero

- El director creativo de Marvel, Kevin Feige, cree que la fase 4 del MCU no ha sido afectada de forma significativa

Año: 2021 Febrero

- Un estudio de la Universidad de Oxford confirma que la inmunización de la vacuna contra la Covid - 19 no debe retrasarse al contrario de lo que opinan algunos expertos
- Los candidatos a la presidencia de la Generalitat podrán disfrutar de un permiso de hasta cuatro horas dentro de la jornada laboral
- El psicólogo Javier Urra destaca que el aumento de los casos de violencia de menores es uno de los principales detonantes de la crisis sanitaria
- La presencia de nuevas variante del virus podría ser en realidad una variante local

Año: 2021 Marzo

- La ministra de Exteriores, Arancha González Laya, asegura que los extranjeros no podrán viajar fuera de su comunidad autónoma
- El 8 de marzo, Día Internacional de la Mujer, estará marcado por la pandemia del coronavirus
- La transformación digital se ha convertido en una herramienta clave para el desarrollo de la transformación digital. Adecco publica un informe sobre los empleos más demandados
- El número de pacientes con Covid - 19 en el primer fin de semana ha aumentado tres centésimas, hasta los 1, 04
- Casi 3, 5 millones de contribuyentes se verán afectados por un ERE en 2020

Año: 2021 Abril

- La Semana Santa de este año vuelve a estar marcada por las restricciones en el marco de la pandemia
- La reina de Inglaterra ha pasado más de 74 años en su casa de Windsor

Año: 2021 Mayo

- El presidente de Ceuta reconoce que la ciudad vive como un " estado de excepción "
- Una selección de las historias de actualidad de la jornada
- Una selección de las historias de actualidad de la jornada

- El número de casos supera los 400. 000, según los datos del Ministerio de Salud indio

Temática “Trump”:

Año: 2020 Enero

- Una selección de las historias de actualidad de la jornada

Año: 2020 Febrero

- El exalcalde de South Bend obtiene 564 delegados en los caucus de Iowa

Año: 2020 Abril

- El presidente de EE UU asegura que 29 de los 50 Estados del país están en condiciones de iniciar la reapertura

Año: 2020 Junio

- Miles de personas recorren las calles de la capital para protestar contra la violencia policial contra los afroamericanos

Año: 2020 Julio

- El rapero y candidato presidencial de Kim Kardashian confirma su candidatura a la presidencia de EE UU

Año: 2020 Agosto

- La aplicación de Trump ofrece 20. 000 millones de dólares al presidente de EE UU

Año: 2020 Octubre

- El presidente de EE UU se encuentra en estado terminal de la enfermedad en el Hospital Walter Reed

Año: 2020 Noviembre

- El jefe de campaña de Trump, Bill Stepien, pide que todos los condados separen cualquier papeleta que hubiera llegado tarde respecto a las elecciones

Año: 2021 Enero

- El presidente electo de EE UU, Joe Biden, se niega a declarar en el primer juicio
- La red social canceló la cuenta del presidente de EE UU tras el asalto al Capitolio en Washington

3. RESULTADOS ALTERNATIVA 3

Temática “pandemia”:

Año: 2020 Marzo

- El número de casos supera los 56. 000 en un día, mientras que las muertes ya han superado el millar
- El número de contagios por Covid - 19 ha aumentado en Hong Kong desde el inicio de la enfermedad
- Anabel Pantoja cancela su boda en cuanto termine la crisis mundial por el aumento de muertos y avanza a peor ritmo
- El presidente del Instituto Superior de Salud, Silvio Brusaferro, asegura que la curva de nuevos casos parece atenuarse ligeramente en su ascenso
- Don Felipe y doña Letizia conversan en el Palacio de la Zarzuela con los ministros de Sanidad y Sanidad
- La Generalitat refuerza la atención asistencial en la Comunidad Valenciana y el resto de España

Año: 2020 Abril

- El Atlético de Madrid, el único equipo español que había sellado su clasificación para los cuartos de final, es el único club español que ha sellado la clasificación para la ChampionsLeague
- La presidenta de la Comunidad de Madrid pide a los presidentes de las comunidades autónomas que faciliten una lista de infraestructuras
- El número de fallecidos en Cataluña supera los 3.756 personas en las 24 horas
- La Consejería de Sanidad contabiliza 3.103 profesionales del Servicio Madrileño de Salud
- La red de mensajería de Facebook cuenta con 1.730 millones de usuarios diarios activos en todo el mundo
- Perla, una mascota mestiza de siete años y medio que lleva unos meses en el centro, ha sido trasladada a Alemania
- La cifra de infectados supera los 33.700, según las cifras oficiales
- King regalará vidas gratuitas e ilimitadas a los usuarios de Candy Crush Saga
- El Hospital San Pedro de Alcántara, de Cáceres, atendió el pasado 12 de marzo a la madre y al bebé
- Las cuatro fases del plan de desescalada del plan para la transición hacia una nueva normalidad en España

Año: 2020 Mayo

- El número de fallecidos por Covid - 19 cae por debajo de la jornada anterior
- El líder del PP cree que la pandemia está demostrando que el Gobierno, "salvo al 8-M, siempre llega tarde a todo"
- El presidente de EE UU reconoce que la cifra de fallecidos en combate será de 100.000, al mismo tiempo que el día anterior
- El ministro del Interior, José Luis Martínez - Almeida, dice que la Guardia Civil no puede rendir cuentas ante el ministro
- El Gobierno de Maduro rechaza las acusaciones del régimen de Nicolás Maduro sobre el estado de Venezuela

Año: 2020 Julio

- El presidente y su ministra de Asuntos Exteriores se embarcan en una ronda europea para lograr que los 140. 000 millones que España calcula que España estima que le corresponderá en las mejores condiciones
- El gigante latinoamericano tiene una incidencia de la enfermedad de 957, 5 casos por cada 100. 000 habitantes
- Arabia Saudí supera el umbral de los 200. 000 casos confirmados
- ¿ Se puede celebrar la junta de manera telemática o por cualquier otro medio que asegure la mayor asistencia posible de los vecinos?

Año: 2020 Septiembre

- El Gobierno vasco ha convocado tres días de huelga para que los niños vuelvan a clase
- El Gobierno, patronal y patronal negocian los ERTE en la semana que viene

Año: 2020 Octubre

- La OMS alerta de que los casos se detectan en comunidades que cumplen poco las medidas de autoprotección
- El Gobierno de Francia amplía el toque de queda a 54 departamentos y la Polinesia
- El riesgo de rebrote es de 1. 756 en la comarca de la Cerdanya (Girona)
- Don Felipe VI, acompañado de la reina Letizia, entregará los Premios Princesa de Asturias 2020

Año: 2020 Noviembre

- El Gobierno de Juanma Moreno ha ampliado dos semanas más el cierre perimetral de toda la comunidad
- Los consejeros de Sanidad de Andalucía, País Vasco y Castilla y León insisten en que el Gobierno central debe tener preparado y planificado un posible confinamiento domiciliario
- El candidato demócrata se muestra confiado de que acabará ganando en los cuatro estados más pobres
- Si tenemos ansiedad será todavía más importante que no tengamos carencias en ninguna de las vitaminas del grupo B

- Sanidad asegura que cada gobierno autonómico deberá presentar un plan para facilitar la detección del virus

Año: 2020 Diciembre

- Una treintena de personas se concentran en el centro de la capital andaluza en protesta por las nuevas medidas de separación
 - El ministro de Sanidad y los consejeros autonómicos volverán a reunirse este miércoles en un Consejo Interterritorial de Sanidad
 - Sanidad prevé inmunizar a dos millones y medio de personas en España
 - El número de personas que se han quedado en el décimo premio del Sorteo de Navidad se ha multiplicado por tres en la última semana
 - El Gobierno de Rajoy estudia cómo compensar a los grandes tenedores de propiedades
 - El fenómeno de las Gemínidas, de unos 2 minutos de duración, se podrá observar en Chile y Argentina
-

Año: 2021 Enero

- En 2020 no se ha estrenado una sola película de Marvel, sino que tanto Godzilla vs. Kong y Dune iban tanto a salas como a HBO Max

Año: 2021 Febrero

- La inmunización de inmunizar a los virus de la inmunización en Reino Unido es clave para evitar el contagio del virus
- La participación en el primer avance a las 13.00 sigue a la baja
- El portavoz de la Comunidad de Madrid en el centro defiende que los padres deben ser conscientes del daño que hacen y se hacen los padres
- La variante brasileña de la variante brasileña se ha detectado en 54 países en España

Año: 2021 Marzo

- La ministra de Asuntos Exteriores, Arancha González Laya, pide a los ciudadanos ser " tremendamente responsables " y abstenerse
- Las mujeres son las grandes afectadas en una crisis sanitaria que echa mano de sus trabajadoras esenciales
- El 72 % de los encuestados cree que el teletrabajo es positivo para las empresas frente a un 49, 9 %
- El número total de fallecidos y 1. 470 en las 24 horas se ha disparado en las últimas 24 horas
- La Agencia Tributaria recomienda a los trabajadores que presenten declaración de IRPF antes de presentar la declaración

Año: 2021 Abril

- La Semana Santa de este año será la primera fiesta popular que se puede celebrar en la calle
- La reina de Inglaterra, de 73 años, permanecerá en la capilla del rey Jorge VI

Año: 2021 Mayo

- La ministra de Defensa, Margarita Robles, destaca el papel de España en la lucha contra la inmigración irregular
- El Consejo Interterritorial en contra del decreto de Rajoy sobre las restricciones a las que las autonomías limitarán los derechos fundamentales
- El 31 % de los españoles prefiere visitar tres o más o más centros
- El país asiático es el segundo país con más muertes en términos absolutos, solo por detrás de Estados Unidos

Temática "Trump":

Año: 2020 Enero

- El primer ministro de EE UU, Justin Trudeau, asegura que los ataques se han saldado sin bajas en el ataque

Año: 2020 Febrero

- El presidente de EE UU se impone con claridad en los caucus de New Hampshire, el favorito para la Casa Blanca

Año: 2020 Abril

- El gobernador de Nueva York, Andrew Cuomo, tiene previsto iniciar la reapertura el 15 de mayo

Año: 2020 Junio

- El presidente de EE UU, Donald Trump, ha pedido al presidente de Estados Unidos que mantenga la boca cerrada

Año: 2020 Julio

- El rapero y candidato presidencial Kim Kardashian y su esposa, la estrella televisiva Kanye West, se reunieron esta semana en el rancho de Wyoming

Año: 2020 Agosto

- El presidente de EE UU asegura que la empresa no podrá seguir presente en EE UU a finales de septiembre

Año: 2020 Octubre

- El presidente de EE UU asegura que la prueba será en los próximos días, pero no aclaró cuándo fue la última vez

Año: 2020 Noviembre

- La jornada electoral en Estados Unidos transcurrió sin grandes incidencias, pero los estadounidenses no saben quién será el ganador de las presidenciales

Año: 2021 Enero

- El primer juicio político al presidente electo de EE UU, que podría ser destituido, se celebrará en la Cámara de Representantes
- La red social suspende la cuenta de Trump ante " el riesgo de una mayor incitación a la violencia "

4. RESULTADOS ALTERNATIVA 4

Temática "pandemia":

Año: 2020 Marzo

- Además, Tedros ha subrayado la importancia de hacer frente al nuevo coronavirus con iniciativas "agresivas y específicas", entre las que ha destacado la realización de la prueba a toda persona sospechosa de padecerlo, asilar y tratar a los casos confirmados y poner en cuarentena a las personas que hayan tenido un contacto estrecho con un infectado. La propagación más rápida del virus está ocurriendo sin duda en Estados Unidos, que está cerca de los 65.000 casos, lo que implica más de 11.000 en un solo día, mientras que las muertes ya han superado el millar.
- En España, el número de contagios ha llegado este miércoles a superar los 2.100 casos y 55 personas han muerto con coronavirus. Así, hasta el viernes, se han registrado un total de 81.008 casos confirmados de coronavirus en la China continental y 3.255 personas han muerto por la enfermedad en el país asiático.
- En este grupo se encuentra precisamente Anabel Pantoja que, desde Gran Canaria, donde ya estrena su nueva casa, ha decidido junto a su novio, Omar Sánchez, que no es el momento, que no se va a poder celebrar y que, en resumidas cuentas, cancelan su boda. Está desesperado.
- Por otra parte, el presidente del Instituto Superior de Salud (ISS), Silvio Brusaferro, ha explicado este viernes por la mañana que desde el 19-20 de marzo se ha constatado que "la curva de nuevos casos parece atenuarse ligeramente en su ascenso", si bien ha recalcado que aún hay zonas con un elevado ritmo de contagios. En este sentido, el comercio minorista ha retomado la actividad este lunes 18 de mayo.
- El rey Felipe VI dirigirá un "mensaje a la nación" el miércoles a las 21.00 horas, después de la reunión que mantendrá en el Palacio de la Zarzuela con el presidente del Gobierno, Pedro Sánchez, y el comité de gestión técnica de la epidemia de coronavirus, ha informado la Casa Real. El Comité Técnico de Gestión del Coronavirus está integrado por el presidente del Gobierno, Pedro Sánchez, y los ministros de Sanidad, Salvador Illa, Defensa, Margarita Robles, Interior, Fernando Grande-Marlaska, y Transportes, Movilidad y Agenda Urbana, José Luis Ábalos.

- El centro se ha reforzado con 20 enfermeros, 34 técnicos sanitarios y dos médicos, que se han unido a su personal. Una en Torrent (Valencia), donde hay 64 personas con síntomas (35 casos confirmados) y cinco pacientes en la UCI.

Año: 2020 Abril

- Con la incertidumbre del parón futbolístico y deportivo debido a la pandemia del coronavirus, a más de uno le tiemblan las piernas pensando que el Atlético pueda perderse la próxima edición de la Champions. El mundo del fútbol continúa paralizado por la pandemia de coronavirus, pero los jugadores de los equipos de LaLiga no pasan el confinamiento de brazos cruzados.

- También ha pedido un presupuesto suficiente para poder afrontar la crisis el presidente de Murcia, Fernando López Miras, quien ha expresado su preocupación por el incremento de movilidad hacia segundas residencias y ha pedido refuerzo con militares y más sanciones. El presidente del Gobierno, Pedro Sánchez, ha pedido este domingo a los presidentes de las comunidades autónomas que faciliten una lista de infraestructuras para poder alojar a contagiados por el coronavirus pero que están asintomáticos, con el fin de evitar que puedan contagiar a otras personas, según han informado fuentes autonómicas.

- En cuanto a las residencias de gente mayor, un total de 10.366 personas han sido confirmadas como positivas de coronavirus y 21.292 son casos sospechosos. En cuanto a las residencias de mayores, un total de 10.556 personas han sido confirmadas como positivas de coronavirus y 22.518 son casos sospechosos.

- En un comunicado remitido este lunes, CESM-CLM ha trasladado su más sentido pésame a la familia, amigos y compañeros del médico fallecido y a todos los profesionales del Sescam y de todo el territorio nacional que han muerto a causa del coronavirus. A ellos se suma el Hospital de Alcalá de Henares con 373 profesionales en aislamiento y contagiados, que es precisamente uno de los centros con mayor presión asistencial por coronavirus.

- Pese a los buenos resultados ofrecidos este miércoles, Facebook dijo que estos solo incluyen una pequeña parte (las tres últimas semanas) del nuevo escenario abierto a raíz de la pandemia de COVID-19, que ha reducido sustancialmente la demanda de publicidad digital y ha hecho caer el precio de los anuncios. Desde que comenzara la crisis sanitaria, Facebook ha estado trabajando agresivamente para eliminar información falsa, bulos y fake news sobre el coronavirus de sus plataformas, es decir, de Facebook, Instagram y WhatsApp.

- Ha decidido acoger a Perla, una perra mestiza de siete años y medio que llevaba unos meses en el centro y que, cuando las autoridades lo permitan, viajará hacia Alemania, donde la librarán a una entidad de adopción con la cual colaboran. No sabía que estamos en CUA CUA-rentena...Tú que lo sí lo sabes... protege a tu familia👨👩👧👦🐶🐱#QuédateEnCasa pic.twitter.com/fSoy4vPMqe Por otra parte, la ausencia de

personas también tiene efecto sobre los animales urbanos, con los que convivimos de manera habitual.

- Las últimas cifras oficiales divulgadas el domingo indican que los fallecidos en hospitales del país por la COVID-19 son ya 20.732, tras sumarse otros 413 en 24 horas, la cifra diaria más baja registrada desde marzo. Siguen hospitalizadas en el país 26.283 personas por Covid-19, de ellas 4.019 en unidades de cuidados intensivos (UCI), aunque en ambos casos los saldos entre admisiones y altas son negativos, con 551 y 188 pacientes menos en un día, respectivamente.

- Su propia experiencia como investigadora confirma que los videojuegos pueden ayudar a niños en momentos puntuales a desarrollar sus destrezas: "Los utilizamos para trabajar la dislexia o la organización espacial con muy buenos resultados en pocas sesiones en comparación con otras técnicas que utilizábamos antes. De este modo, el Congreso Internacional de Videojuegos y Ocio Interactivo que se viene celebrando en Barcelona desde 2011 no cambia de fecha su cita anual y se adapta a la pandemia permitiendo a todo el mundo acceder a la feria desde cualquier parte del mundo sin tener que desplazarse hasta la ciudad con dal.

- El protocolo que se sigue en estos casos, han explicado las mismas fuentes, es conjunto entre ambos servicios y se aplica a las mujeres que dan positivo en Covid-19, con un seguimiento mayor durante el embarazo y se planifica el parto. El primer día que el Puerta de Hierro puso en marcha este protocolo, de siete mujeres que acudieron al hospital a dar a luz, cuatro dieron positivo en coronavirus.

- El plan de desescalada que ha anunciado el Gobierno este martes, consta de cuatro fases diferenciadas, en cada una de las cuales se permitirá ir realizando más actividades, siempre con medidas de seguridad e higiene, hasta llegar a lo que Pedro Sánchez denominó en la rueda de prensa como "nueva normalidad". Las cuatro fases del plan de desescalada que ha elaborado el Gobierno permite, según la fase en la que se encuentre cada zona del territorio español, ir realizando cada vez más actividades y abriendo, en consecuencia, locales, comercios y negocios.

Año: 2020 Mayo

- En los últimos 14 días, 995 personas han iniciado síntomas de la enfermedad, de los que 282 los iniciaron durante la última semana. En los últimos 14 días, 1.071 personas han iniciado síntomas de la enfermedad, de los que 320 los iniciaron durante la última semana.

- Ya habrá tiempo de que haya pelea política en el Congreso", ha insistido Iglesias, quien si bien ha dicho entender que el PP "utilice" Madrid y aquellas comunidades en las que gobierna para "atacar" al Gobierno, "pero no para atacar a los epidemiólogos, a los especialistas y a los profesionales de la sanidad. Hay un concepto que sobrevuela, que habría que recuperar, que es el de nacionalizar.

- El balance provisional de fallecidos -91.845- sigue por debajo de las estimaciones iniciales de la Casa Blanca, que proyectó en el mejor de los casos entre 100.000 y 240.000 muertes; pero ha superado ya con creces los cálculos más optimistas que hizo "a posteriori" el presidente estadounidense, Donald Trump, de entre 50.000 y 60.000 fallecidos. El balance provisional de fallecidos -84.059- sigue por debajo de las estimaciones iniciales de la Casa Blanca, que proyectó en el mejor de los casos entre 100.000 y 240.000 muertes; pero ha superado ya con creces los cálculos más optimistas que hizo "a posteriori" el presidente Donald Trump de entre 50.000 y 60.000 muertos.

- El ministro del Interior, Fernando Grande-Marlaska, ha desvinculado el cese del coronel jefe de la Comandancia de la Guardia Civil en Madrid, Diego Pérez de los Cobos, del informe sobre las manifestaciones del 8M que el instituto armado remitió a un juzgado. Laurentino Ceña Coro, hasta ahora número dos de la Guardia Civil, ha dimitido este martes después de que el ministro de Interior, Fernando Grande Marlaska, cesara al jefe de la Comandancia del instituto armado en Madrid, el coronel Diego Pérez de los Cobos, por "pérdida de confianza".

- Según la Cancillería, las afirmaciones del "régimen dictatorial de Nicolás Maduro", culpando a Colombia "de supuestos hechos de desestabilización", son un intento más de "desviar la atención respecto de los verdaderos problemas que vive el pueblo de Venezuela". Después de la orden judicial de este martes, el opositor se queda sin liderazgo parlamentario y, por tanto, sin un respaldo político para erigirse en presidente encargado de Venezuela frente a sus seguidores y la comunidad internacional, como lo ha hecho hasta la fecha.

Año: 2020 Julio

- El objetivo es salir de la crisis con una Europa más fuerte. Los llamamientos al acuerdo y a la responsabilidad de algunos líderes, entre ellos el presidente español Pedro Sanchez, quedaron eclipsados con la realidad que puso encima de la mesa la canciller alemana, Angela Merkel, a su llegada al Consejo Europeo: "espero unas negociaciones muy difíciles".

- Brasil registró en las últimas 24 horas 39.924 nuevos casos de coronavirus, con lo que ya bordea los 2 millones de contagios, así como 1.233 nuevas muertes, con lo que superó los 75.000 fallecidos por la pandemia, informó este miércoles el Ministerio de Salud. Con el promedio de cerca de 35.000 nuevos casos diarios en los últimos días, es muy probable que Brasil alcance el jueves los 2 millones de contagiados, que lo confirman como uno de los nuevos epicentros globales de la pandemia y el segundo país en el mundo con más infectados y muertes después de Estados Unidos.

- Por su parte, Colombia registra 276.055 personas contagiadas y 9.454 fallecidos, seguida por Arabia Saudí (272.590 casos y 2.816 muertos), Italia (246.776 positivos y 35.129 decesos), Bangladesh (232.194 contagios y 3.035 fallecidos), Turquía (228.924 casos y 5.659 decesos), Francia (221.077 positivos y 30.226 fallecidos) y Alemania (208.546 positivos y 9.145 muertos) cierran la lista de los países que super

an el umbral de los 200.000 casos confirmados. Arabia Saudí supera ampliamente el umbral de los 250.000 casos, con un total de 258.156 personas contagiadas y 2.601 fallecidos, seguida por Italia (245.032 casos y 35.082 muertos), Turquía (222.402 casos y 5.545 decesos), Colombia (218.428 contagios y 7.373 fallecidos), Francia (215.605 casos y 30.175 decesos), Bangladesh (213.254 positivos y 2.751 fallecidos) y Alemania (204.480 positivos y 9.102 muertos) cierran la lista de los países que superan el umbral de los 200.000 casos confirmados.

- Él es una persona muy cariñosa y desde que nos vimos después del confinamiento ha estado frío y distanciado. Tanto si la encuentra como si no, es necesario que se perdone a usted mismo y a la persona que le ha hecho daño.

Año: 2020 Septiembre

- Además, reclaman al Gobierno de Unidas Podemos que "den un puñetazo en la mesa del Consejo de Ministros", que no acepten la situación y que atiendan "las peticiones justas" de la comunidad educativa, de los profesores y los estudiantes El primero de los tres días de huelga convocado por el Sindicato de Estudiantes (SE) en los centros educativos para exigir unas "aulas seguras" ha tenido un seguimiento desigual mientras que las incidencias por el coronavirus se suceden y ya suman más de 330, entre cierre de aulas y colegios. El Sindicato de Estudiantes (SE), que ha convocado tres días de huelga en los centros educativos para demandar un regreso seguro a las aulas, ha tenido que suspender las concentraciones previstas para este jueves en distintos puntos de España ante el avance de los contagios de covid.

- Una de las grandes diferencias entre la patronal y el Gobierno, que ya se produjo en la anterior negociación, tiene que ver con el diseño de las exenciones fiscales para las empresas en los ERTE. La semana que viene, Ejecutivo, patronal y sindicatos tendrán también que intentar desencallar otros dos asuntos que están dilatando la negociación: el formato de las exoneraciones fiscales para las empresas acogidas a un expediente y los sectores que podrán seguir utilizando los ERTE por fuerza mayor.

Año: 2020 Octubre

- No obstante, ha subrayado la importancia de que los gobiernos ayuden a sus poblaciones a poder cumplir con las medidas impuestas. Desde la pasada semana entró en vigor la prohibición de viajar a Gales para aquellas personas que procedan de lugares del Reino Unido en los que se registran altos niveles de infección.

- Este miércoles se han registrado 22.591 nuevos contagios y la presión hospitalaria ha aumentado en gran parte del país. Francia ha confirmado este viernes su cifra récord de nuevos contagios de la Covid-19, más de 20.000 en las últimas 24 horas, según los datos difundidos por el Ministerio de Salud francés.

- Respecto a la comarca de la Cerdanya (Girona), desde el inicio de la pandemia ha habido 491 casos confirmados acumulados de coronavirus, de los que 6 han muerto, mientras que actualmente hay 5 pacientes ingresados --2 de ellos en la UCI--; el riesgo de rebrote es de 356,88. Respecto a la comarca de la Cerdanya (Girona), desde el inicio de la pandemia ha habido 460 casos confirmados acumulados de coronavirus, de los que 6 han muerto, mientras que actualmente hay 2 pacientes ingresados -ninguno de ellos en la UCI-; el riesgo de rebrote es de 417,69.

- La ceremonia, a la que como todos los años asiste la reina Sofía y en la que por segundo vez toma la palabra la princesa de Asturias antes de que lo haga Felipe VI, se celebra en la antigua capilla del Hotel Reconquista con un aforo limitado a autoridades y la asistencia de premiados en cinco de las ocho categorías. La ceremonia, a la que como todos los años asistirá la reina Sofía y en la que por segundo año tomará la palabra la princesa de Asturias justo antes de que lo haga el Rey Felipe, se celebrará en la antigua capilla del Hotel Reconquista con un aforo limitado a autoridades y la asistencia de premiados en cinco de las ocho categorías.

Año: 2020 Noviembre

- Moreno ha recordado que la provincia de Granada mantiene el resto de limitaciones vigente en Andalucía, el toque de queda entre las 22.00 y las 7.00 horas y el cierre perimetral de todos los municipios, que se sitúan en nivel 4 de alerta grado 1, "una circunstancia preocupante ante la que no podemos relajarnos", ya que ha advertido de que todo lo avanzado se puede perder en 14 días. Además, los positivos contabilizados esta jornada son 1.488 menos que el domingo de la semana.

- Los consejeros de Sanidad de Andalucía, País Vasco y Castilla y León han insistido este miércoles en el Consejo Interterritorial de Sanidad en que el Gobierno central debe tener preparado un nuevo confinamiento domiciliario si la situación de la pandemia sigue siendo grave y las medidas que han puesto en marcha los gobiernos autonómicos no surten efecto. Un sentir muy diferente es el que se vive en la Puerta del Sol, sede del Gobierno de la Comunidad de Madrid, que no contempla una situación como la de la pasada primavera.

- Este martes, 3 de noviembre, el candidato Donald Trump ha sorprendido con un electorado fiel que le ha permitido librar una apretada batalla contra el aspirante demócrata, Joe Biden, a quien las encuestas anticipaban, de nuevo erróneamente, un camino fácil hacia la Casa Blanca. Por ello, el candidato demócrata se ha mostrado confiado de que acabará ganando en estos cuatro estados, lo que le daría un amplio margen respecto a Trump.

- De esta manera, hay que evitar la sobreinformación y el estar permanentemente conectados porque podría "aumentar su sensación de riesgo y nerviosismo". En la primera ola teníamos sensación de miedo e incertidumbre.

- Aunque esta recomendación no es vinculante, este jueves, la ministra de Industria, Comercio y Turismo, Reyes Maroto, dijo en televisión que confiaba en poder empezar a usar pronto también test de antígenos en los "corredores turísticos seguros", comunidades que realizan controles a la entrada y a la salida de los viajeros. La Comunidad de Madrid, por su parte, también tiene puesta la mirada en la validación un protocolo para los test de antígenos de cara a la Navidad.

Año: 2020 Diciembre

- Agentes de la Policía Local de Sevilla desalojaron este martes por la noche a una treintena de personas que celebraba una fiesta de cumpleaños en una barriada de la capital andaluza, sin respetar las medidas destinadas a frenar la pandemia de COVID-19, informa Efe. Unos 300 trabajadores del sector de la hostelería de Granada se han concentrado este sábado por la tarde en el centro de la capital para protestar por las nuevas medidas decretadas por la Junta para frenar la pandemia de la covid-19.

- Estas son las dos nuevas circunstancias en las que el ministro de Sanidad, Salvador Illa, y los consejeros autonómicos volverán a reunirse este miércoles en un Consejo Interterritorial de Sanidad que previsiblemente se revisarán las fechas, para adelantarlas, de la llegada de las primeras dosis de vacuna, así como el Plan de Navidad, que algunas comunidades ya están reforzando ante el nuevo repunte de la pandemia. En todo caso, ha asegurado que se seguirán recibiendo de manera cadencial y semanalmente, lo que hará que en poco tiempo se mezclen con la segunda vacuna que está previsto que autorice la Comisión Europea, la de Moderna, que será analizada por la Agencia Europea del Medicamento (EMA) el 8 de enero.

- Es más, una amplia mayoría (más del 82%) de los españoles confirman que se han estado informando de los últimos avances en lo que a la vacuna se refiere. En lo que se refiere a la responsabilidad individual, el 50,6% de los españoles apuntan que la ciudadanía está siendo responsable ante el virus, pero hay un 37,7% que opina lo contrario: no hay civismo en la situación en la que nos encontramos.

- Son muchas las opciones de envío con las que contamos para que el regalo llegue a su destinatario. Si el premio es superior a 2.500 euros, deberá llevar el/los décimos a cualquiera de las entidades bancarias autorizadas por Loterías y Apuestas del Estado, a partir del día siguiente.

- Con el nuevo decreto, previsto que se apruebe antes de Navidad, los Servicios Sociales deberán hacer un informe que, si dictamina que el afectado entra en la categoría de vulnerable, obligará a buscar una solución habitacional en un máximo de tres meses. Y apuntan que se están estudiando con la Vicepresidencia Segunda fórmulas para que los servicios sociales puedan gestionar este descuento a potenciales beneficiarios que no saben que pueden disfrutarlo. Esta es la segunda vez en apenas dos meses que PSOE y Unidas Podemos se enfrentan a raíz de la negativa de los socialistas a prohibir el corte de los suministros básicos a familias vulnerables.

- El máximo de la lluvia de las Gemínidas se producirá durante las noches del 12 y 13 de diciembre, mientras que en la madrugada del 14 de este mes en territorios de Chile y Argentina se podrá observar un eclipse total de Sol de unos 2 minutos de duración, y siete días después Júpiter y Saturno estarán a la distancia mínima entre ellos. Pese a ser uno de los epicentros de la pandemia en Chile durante las últimas semanas, más de 70.000 turistas han llegado a la Región de La Araucanía en la zona sur del país para observar el eclipse solar que este lunes se podrá apreciar en su totalidad desde ese punto.

Año: 2021 Enero

- La novedad más grave, claro, tiene que ver con los grandes blockbusters de Sony. ¿Cuál es vuestra historia?

Año: 2021 Febrero

- La cuestión, tal y como admite por ejemplo el Gobierno británico, es que aún no hay estudios que revelen cuántos vacunados se infectan y pueden transmitir el coronavirus, aun no teniendo síntomas. "Los coronavirus son menos propensos a la mutación que los virus de la gripe, pero siempre hemos previsto que, a medida que la pandemia continúe, las nuevas variantes comenzarán a ser dominantes entre los virus que están circulando y que eventualmente se requerirá una nueva versión de la vacuna, con una proteína de pico actualizada, para mantener la eficacia de la vacuna al nivel más alto posible", detalla Sarah Gilbert, catedrática de Vacunología e investigadora principal del ensayo de la vacuna de Oxford.

- En cuanto al detalle de los resultados el candidato de los socialistas catalanes, Salvador Illa, obtuvo el 22,99 % de los votos (645.892) y -con el 99,10% escrutado- colocó al PSC como la fuerza más apoyada en Cataluña, seguido muy de cerca por ERC con el 21,32% (598.921). La participación en el segundo avance a las 18.00 de la tarde continuaba siendo baja y se situaba en un 45,72% del censo electoral.

- Para Royo el sistema actual tampoco pone de su parte para solucionar la situación y para que los jóvenes sean conscientes del daño que hacen y se hacen: "Cuando un adolescente es multado por organizar un botellón, pelearse con las fuerzas policiales o saltarse el toque de queda, son los padres, en la gran mayoría de los casos, los que deben hacer frente a ese pago en nombre de sus hijos. Y hacer esto tampoco está mal porque, sinceramente, en todas las familias hay cosas cuestionables, los padres también deben estar abiertos al escrutinio, se pueden revisar las nociones que a veces son muy firmes en las familias a raíz de la contribución de un adolescente.

- Afortunadamente, el virus muta relativamente poco y eso podría significar que las primeras vacunas tolerarán estas primeras variantes que se van asentando. "Vimos que para estudiar a fondo cualquier virus, lo interesante es hacer secuenciaciones y conocer su estructura y composición.

Año: 2021 Marzo

- Pero además de la llegada de turistas alemanes a España, también están generando polémica los viajes de turistas franceses, en su mayoría jóvenes, que desde hace semanas aterrizan en masa en Madrid para disfrutar de los bares y el ocio de la capital española, unas 'libertades' que actualmente no encuentran en su país. Estamos tratando de seguir bajando la incidencia.

- La pandemia no trata a todos por igual y, una vez más, las mujeres son las grandes perjudicadas en una crisis sanitaria que echa mano de sus trabajadoras esenciales. No hubo manifestaciones masivas como otros años por culpa de la pandemia, pero el Día Internacional de la Mujer se celebró este lunes en toda España con múltiples movilizaciones feministas repartidas por las principales ciudades, incluida Madrid a pesar la prohibición expresa que pesaba en la capital, corroborada este lunes mismo por el Tribunal Constitucional.

- Entre los aspectos más positivos del teletrabajo para las empresas, los encuestados destacan que se reducen costes (65,9%), se evitan desplazamientos (61,2%) y se aumenta la productividad de los empleados (48,7%); mientras que entre las notas negativas destacan que se aísla a las personas (61,8%), es difícil de controlar (45,4%) y perjudica el trabajo en equipo (42,1%). De hecho, el 72% cree que el teletrabajo permite hacer los descansos necesarios.

- Desde el inicio de la pandemia, en Cataluña se han confirmado por todo tipo de pruebas un total de 582.446 positivos -de ellos 1.470 en las últimas 24 horas-, mientras que el coronavirus ha causado la muerte a un total de 21.113 personas, las últimas 25 notificadas este viernes. La pandemia de coronavirus no da tregua y los datos epidemiológicos siguen al alza, aunque de forma moderada, en Cataluña, con 1.207 contagios y 15 fallecidos las últimas 24 horas, en el primer fin de semana sin confinamiento comarcal, que ha permitido a los catalanes salir más allá de sus lugares de residencia.

- De esta manera, aunque esta prestación esté exenta y el perceptor no haya obtenido ninguna otra renta "deberá presentar declaración de IRPF 2020 y, en este caso, las casillas de su declaración aparecerán con importe cero". El primer paso para presentar el borrador por Internet es acceder a la web oficial de la Agencia Tributaria, concretamente al apartado denominado Renta 2020, y después entrar en el servicio de tramitación del borrador/declaración.

Año: 2021 Abril

- El jueves 1 de abril, a las 10.00 horas, tendrá lugar la Santa Misa a Crismal, que da inicio al Triduo Pascual y estará oficiada por el papa Francisco desde el Vaticano. Aunque la Semana Santa de este año 2021 vuelve a estar marcada por las restricciones en el marco de la pandemia, se pueden seguir las misas y demás actos litúrgicos del Jueves Santo y del Viernes Santo a través de las redes sociales, la televisión o las webs de las diferentes entidades religiosas.

- Aunque el ataúd del príncipe Felipe, consorte de la reina Isabel II durante 73 años, yacerá inicialmente en esa cripta, está dispuesto que cuando la monarca británica muera, se le trasladará a la capilla conmemorativa del rey Jorge VI de la iglesia gótica para que el matrimonio esté enterrado en el mismo lugar. Con la muerte del príncipe Felipe, su confidente y asesor, Isabel II continúa sola los últimos años de su reinado, uno de los más importantes y el más largo de la historia del Reino Unido.

Año: 2021 Mayo

- Robles alzó el tono para trasladar a Marruecos que "cuando se utilizan menores como un instrumento para burlar las fronteras territoriales de España no se puede aceptar desde el punto de vista del derecho internacional y humanitario". Acabo de hablar con el presidente de Ceuta para apoyarle en la grave crisis migratoria que sufre nuestra ciudad autónoma. El Gobierno de España debe garantizar de inmediato la integridad de nuestras fronteras y coordinar con Marruecos la devolución de los inmigrantes a su país.

- El presidente del Gobierno, Pedro Sánchez, afirmó este martes que el estado de alarma es el "pasado" e insistió en la idea de que las comunidades autónomas tienen suficientes herramientas para controlar la pandemia en ausencia de ese paraguas jurídico, que decayó el pasado domingo, y pese a que en las últimas horas se han producido resoluciones contradictorias por parte de distintos tribunales superiores autonómicos. En octubre, y tras un verano convulso también en los tribunales, el Gobierno volvió a aprobar la declaración de un estado de alarma en el que establecía el toque de queda y compartía la cogobernanza con las autonomías.

- Las programaciones de todos ellos se han resentido; si los grandes museos más populares han sentido con dureza este impacto, los menores afrontan una crisis que puede ser letal. "Ha surgido un proceso de autoconcienciación en el que los museos hemos asumido que hay que trabajar mucho para satisfacer a nuestras comunidades y al público local", reconoce.

- El segundo país más afectado del mundo en términos absolutos por la pandemia, solo por detrás de Estados Unidos (32,4 millones), se encuentra sumido en una vertiginosa segunda ola que ha puesto al límite a su sistema de salud, con escasez de oxígeno y camas en grandes ciudades como Nueva Delhi. Después de superar este sábado los 400.000 casos diarios por primera vez, los nuevos contagios se redujeron levemente hasta los 392.488, sumando más de 19,5 millones de positivos desde

el inicio de la pandemia, según los datos del Ministerio de Salud indio.

Temática "Trump":

Año: 2020 Enero

- El acontecimiento clave fue el asesinato, a manos de Estados Unidos, de Qasem Soleimani, líder militar iraní y uno de los hombres más importantes del régimen. Trump aseguró horas después que los ataques se han saldado sin bajas, si bien un responsable de la Guardia Revolucionaria de Irán afirmó que en los mismos han muerto al menos a 80 militares estadounidenses.

Año: 2020 Febrero

- Cuarta clasificada en las encuestas de Iowa y tercera en las nacionales, ha luchado toda la campaña por desmarcarse de Sanders, al que acusa de haberle asegurado en privado que una mujer no podía ser presidenta de EE UU. Durante los tres días que ha durado el recuento tanto Buttigieg -la misma noche electoral con un 0% escrutado- como Sanders este mismo jueves se han declarado ganadores de estas primarias.

Año: 2020 Abril

- Este viernes, en una rueda de prensa, Trump ha pedido esperar a ver el desarrollo del virus en las próximas semanas y ha prometido que escuchará "con mucha atención" a sus asesores en temas de salud pública, aunque ha vuelto a incidir en las ventajas económicas de acabar con el parón económico cuanto antes. El estado de confinamiento en los hogares por el coronavirus ha hecho que las casas demanden más papel higiénico de lo habitual hasta el punto de que ya es un bien escaso en los supermercados.

Año: 2020 Junio

- Las protestas no cesan en Estados Unidos, donde al menos 40 ciudades han decretado el toque de queda y permanecen bajo custodia de la Guardia Nacional, e incluso este martes la capital estadounidense vivió un inesperado cacerolazo mientras los manifestantes continuaban frente a la Casa Blanca. Las protestas se extendieron por todo el país con enfrentamientos muy violentos entre la policía y los manifestantes, reivindicando los derechos de la comunidad afroamericana.

Año: 2020 Julio

- El rapero y candidato presidencial Kanye West pidió disculpas a su esposa, la estrella televisiva Kim Kardashian, por revelar detalles privados de su familia en su primer acto de campaña hace una semana, según informaron este domingo medios locales. Justo horas antes, West publicó un tuit en el que aseguraba que estaba intentando divorciarse de su mujer pero que borró poco después.

Año: 2020 Agosto

- Lo que parece claro es que la clave de cualquier acuerdo que una empresa pueda tener con TikTok son los datos y los usuarios a los que se tendrá acceso. Este sábado, Donald Trump anunció que iba a prohibir la aplicación de TikTok en Estados Unidos.

Año: 2020 Octubre

- En ese primer anuncio, Trump ya había explicado que se habían hecho la prueba tras conocer el positivo de la asesora, pues pasaban mucho tiempo trabajando juntos dentro de la campaña presidencial. Lo he aprendido y entendido y os lo voy a contar", aseguró Trump, como si su estancia en el hospital estuviera llegando a su fin.

Año: 2020 Noviembre

- Según las últimas proyecciones de los medios todavía Joe Biden tiene más posibilidades que Trump en este estado, pero no es tan sólida la diferencia como para anotarse ya el resultado. Trump, por su parte, cuenta con un total de 214 delegados y este jueves en una rueda de prensa en la Casa Blanca reforzó su desafío al proceso electoral al cuestionar sin aportar pruebas la legitimidad de millones de votos emitidos por correo, mientras se estrechaban sus opciones de reelección y su campaña libraba litigios en varios estados clave.

Año: 2021 Enero

- El comunicado se ha publicado coincidiendo con la presentación de los demócratas en la Cámara de Representantes este lunes de una solicitud formal para abrir un juicio político o impeachment contra el presidente Trump, alegando también "incitación a la insurrección" por su apoyo a las movilizaciones que culminaron con el asalto al Capitolio

por miles de sus simpatizantes el pasado 6 de enero. Un total de diez legisladores republicanos votaron este miércoles a favor de iniciar un juicio político (impeachment) contra el presidente saliente de EE UU, el también republicano Donald Trump, acusado de "incitación a la insurrección" por su papel en el asalto al Capitolio que llevaron a cabo los partidarios del mandatario la semana pasada.

- Asimismo, es el hogar de quienes rechazan la moderación de contenidos en los canales de Social Media y, ahora, también de los acérrimos defensores de Trump. YouTube se ha unido a otras plataformas y redes sociales como Twitter, Instagram o Facebook y ha decidido suspender el canal del aún presidente de EE UU Donald Trump.