Eric Morales

SISTEMAS DE RECOMENDACIÓN DE NOTICIAS BASADOS EN APRENDIZAJE PROFUNDO

Escuela Politécnica Superior

Ingeniería Informática

Trabajo de Fin de Grado

UAM
UNIVERSIDAD AUTONOMA
DE MADRID

Escuela Politécnica Superior

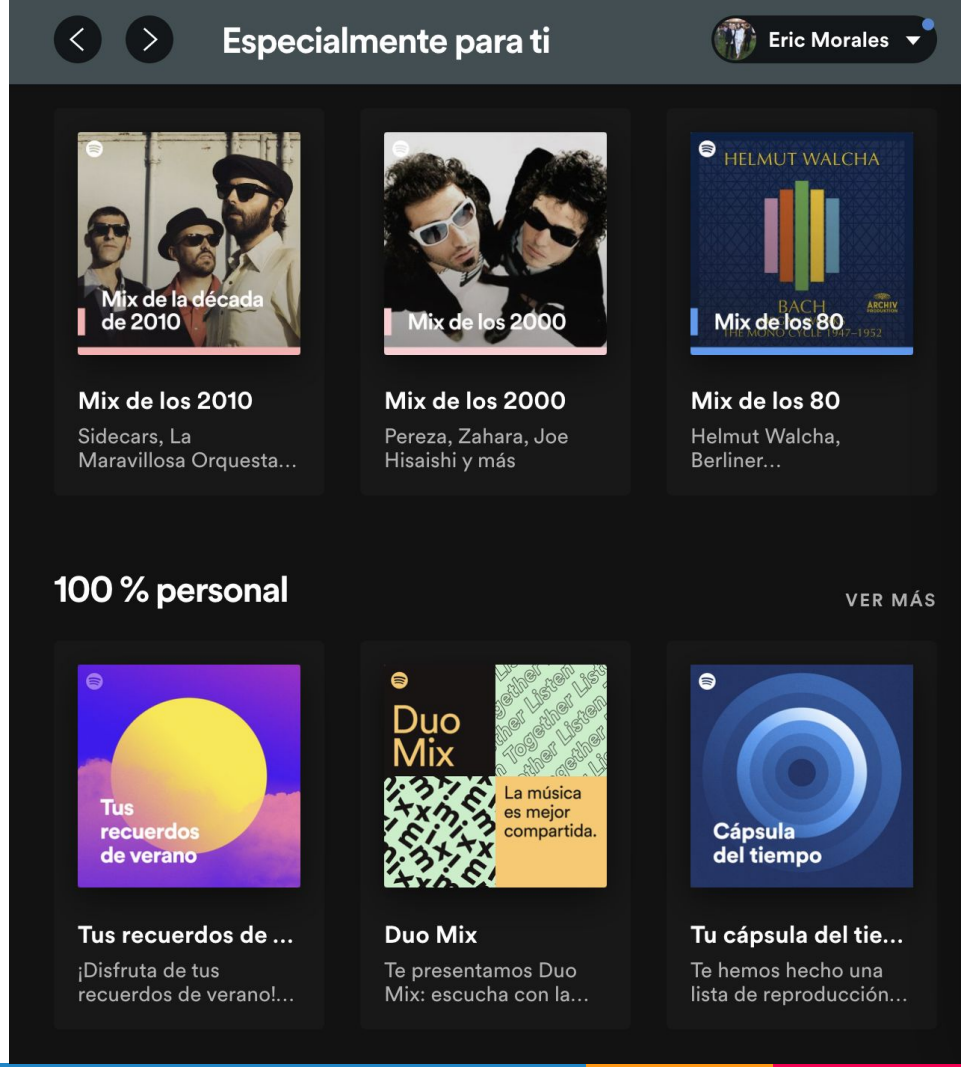
1.

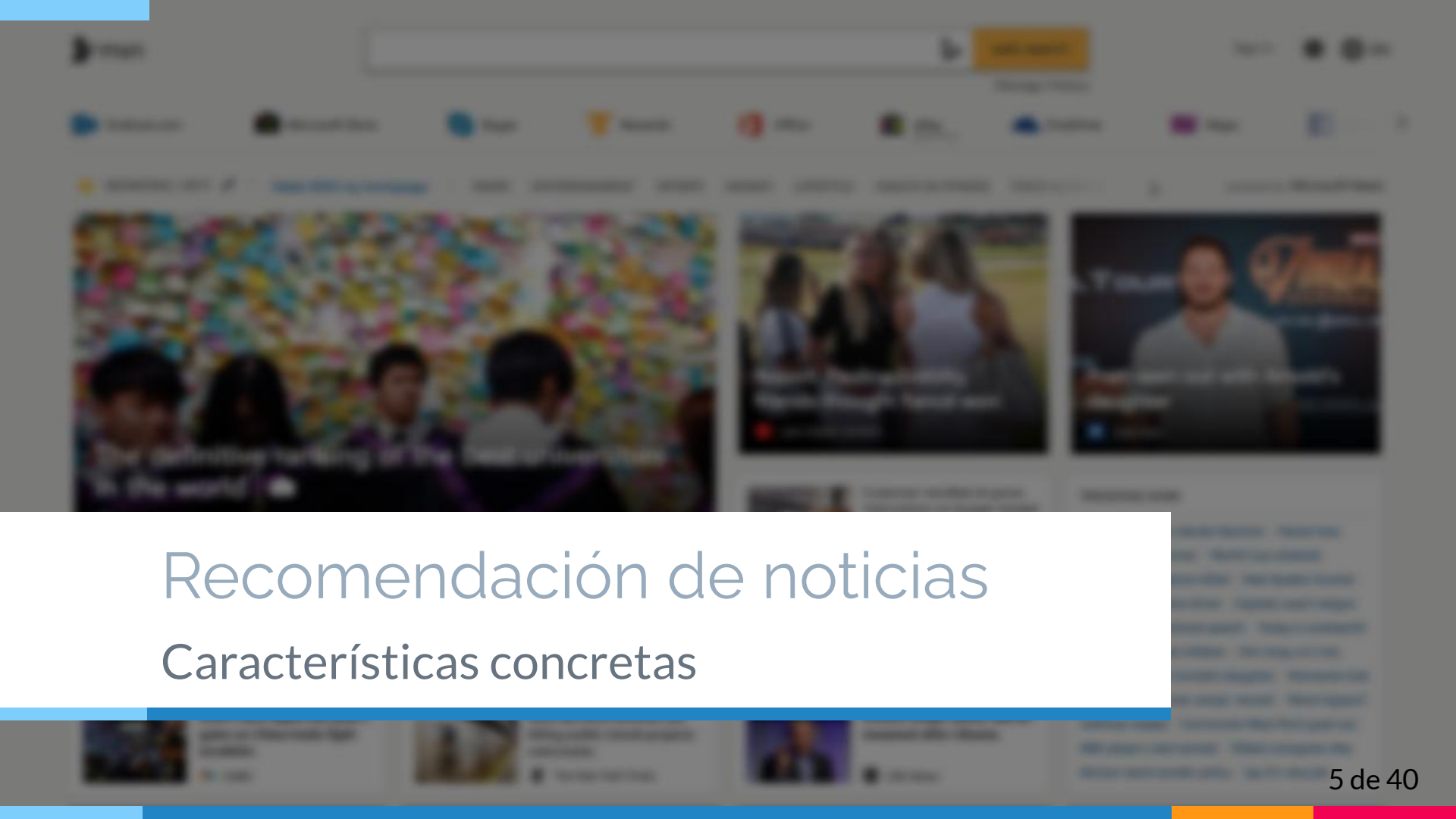
SISTEMAS DE RECOMENDACIÓN

¿Por qué nos interesan?

Sistemas de recomendación

- Sobrecarga de información imposible de consumir.
- Facilitan al usuario la tarea de encontrar ítems afines a sus intereses.





Recomendación de noticias

Características concretas

msn

web search

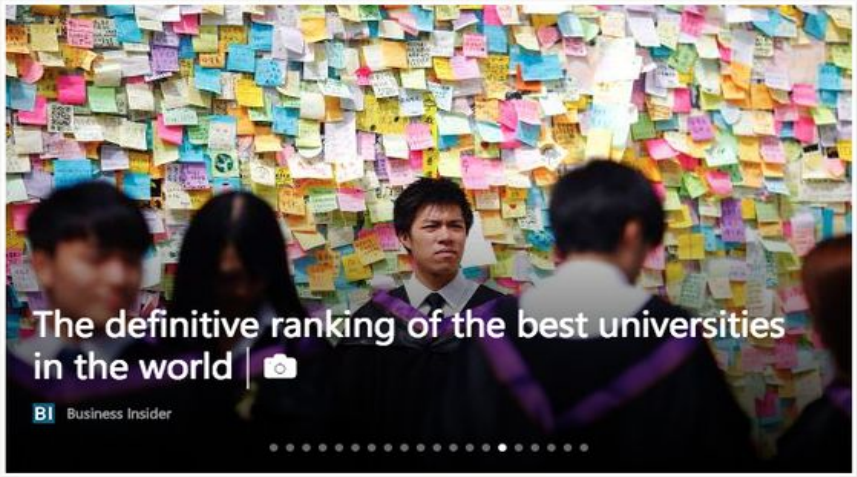
Sign in EN

Manage History

Outlook.com Microsoft Store Skype Rewards Office eBay Sponsored OneDrive Maps

REDMOND / 65°F | Make MSN my homepage | NEWS ENTERTAINMENT SPORTS MONEY LIFESTYLE HEALTH & FITNESS FOOD & DRINK

powered by Microsoft News



Customer revolted at gross instructions on burger receipt

FOX News

Grieving family heartbroken when deceased student's picture left out of yearbook

CNN

Macron admonishes teen, goes viral

Associated Press

VW names interim Audi boss, seeking stability after CEO arrest

Reuters

Dow's dive wipes out year's gains as China trade fight escalates

CNBC

How the Koch brothers are killing public transit projects nationwide

The New York Times

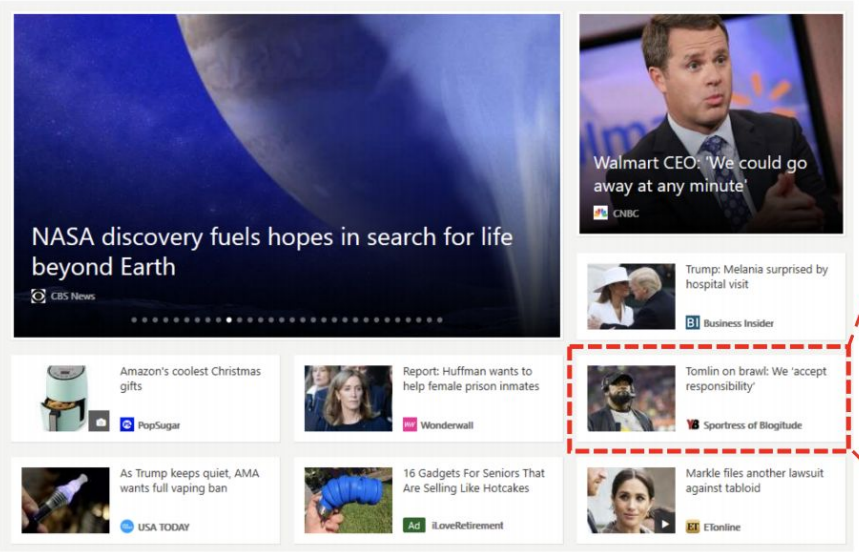
School scraps name, will be renamed after Obama

CBS News

TRENDING NOW

Church members rebuke Sessions | Hawaii lava
Japan wins in stunner | World Cup schedule
Rapper XXXTentacion killed | Kate Spade's funeral
Cop pulls over slow driver | Capitals coach resigns
ESPN host's emotional speech | Today is Juneteenth
Audio of detained children | Kim Jong-un's trip
Pratt picnics with Arnold's daughter | Monsanto trial
Ingraham's 'summer camps' remark | Worst tippers?
Hoffman traded | Communist West Point grad out
NBA player's dad wanted | Oldest orangutan dies
McCain slams border policy | Jay-Z's new job

Recomendación de noticias



(a) An example Microsoft News homepage

Title	Mike Tomlin: Steelers 'accept responsibility' for role in brawl with Browns
Category	Sports
Abstract	Mike Tomlin has admitted that the Pittsburgh Steelers played a role in the brawl with the Cleveland Browns last week, and on Tuesday he accepted responsibility for it on behalf of the organization.
Body	Tomlin opened his weekly news conference by addressing the issue head on. "It was ugly," said Tomlin, who had refused to take any questions about the incident directly after the game, per Brooke Pryor of ESPN. "It was ugly for the game of football. I think all of us that are involved in the game, particularly at this level, …

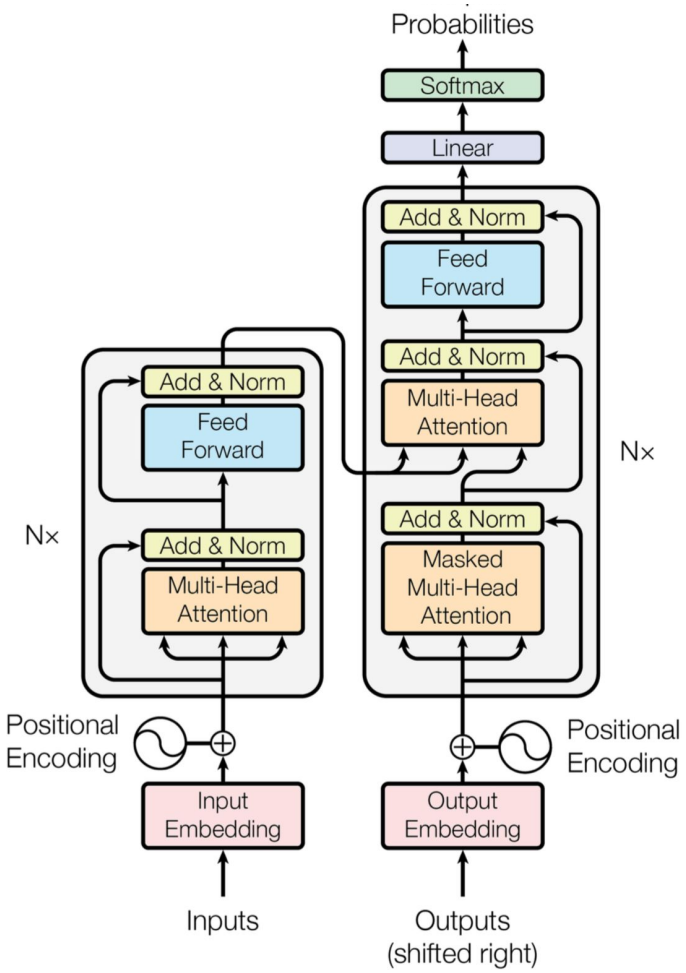
(b) Texts in an example news article

Procesamiento de lenguaje natural

- Las noticias tienen gran parte de su información en forma de texto libre.
- Reto: conseguir que un ordenador comprenda esa información.

Aprendizaje Profundo

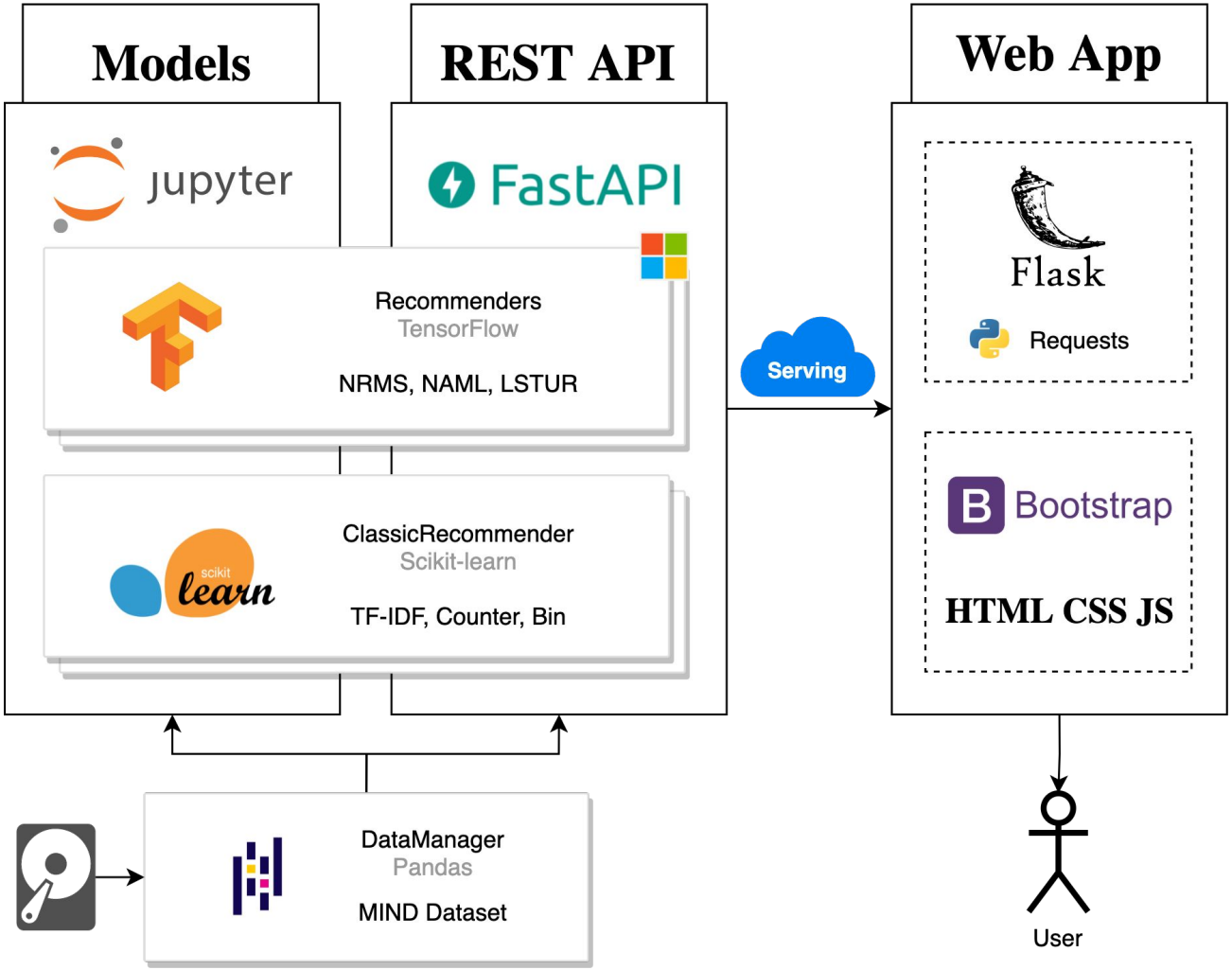
- Han adquirido mucha relevancia en el ámbito del NLP.
- Google con la estructura Transformer incorpora el concepto de self-atención.

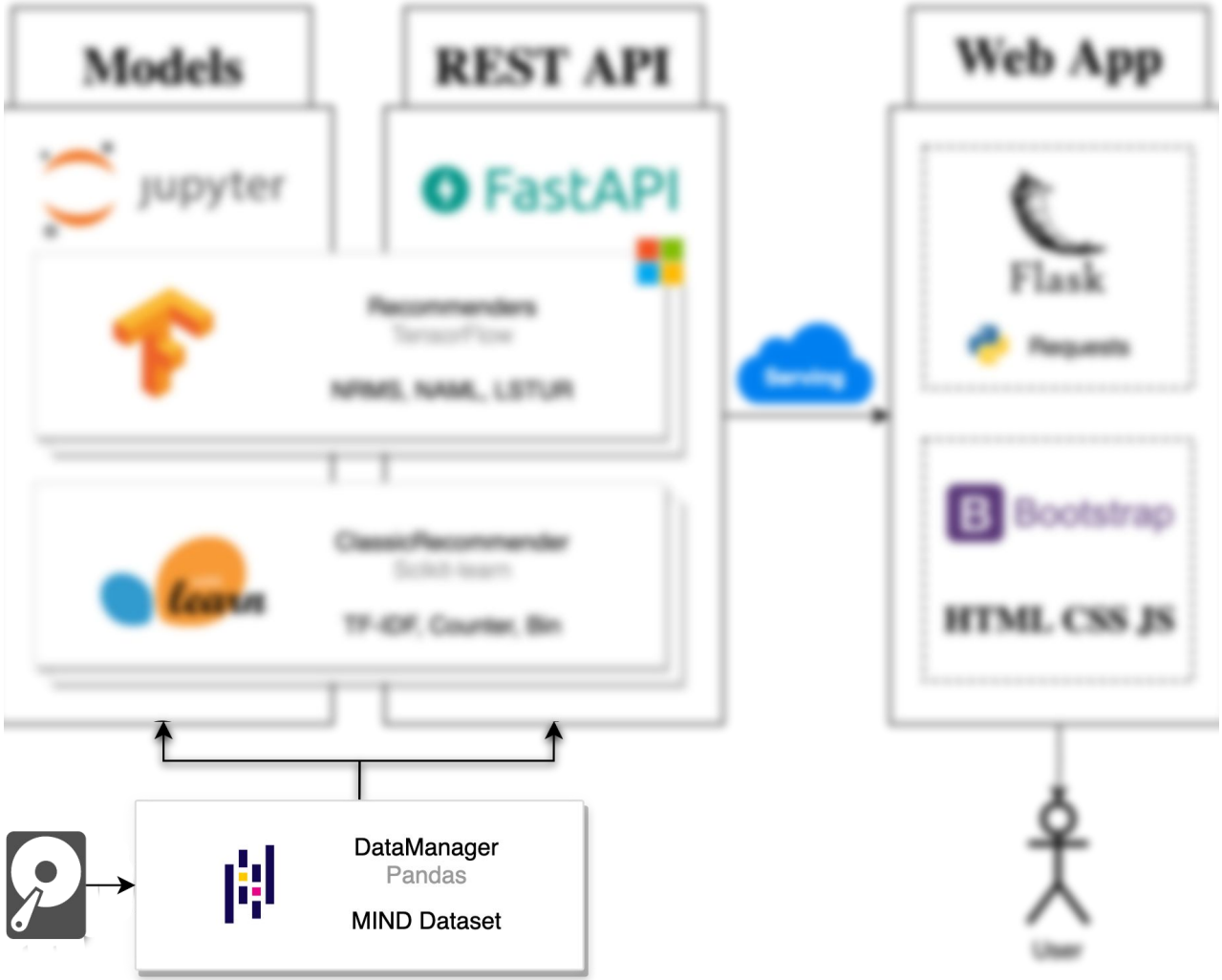


Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

2.
ESTRUCTURA GENERAL

División en módulos del proyecto


Models
REST API
Web App
jupyter
FastAPI
Recommenders
TensorFlow
NRMS, NAML, LSTUR
ClassicRecommender
Scikit-learn
TF-IDF, Counter, Bin
Serving
Flask
Requests
Bootstrap
HTML CSS JS
DataManager
Pandas
MIND Dataset
User


Models
REST API
Web App
jupyter
FastAPI
Recommenders
TensorFlow
NRMS, NAML, LSTUR
Serving
Flask
Requests
Bootstrap
HTML CSS JS
ClassicRecommender
Scikit-learn
TF-IDF, Counter, Bin
DataManager
Pandas
MIND Dataset
User

MIcrosoft News Dataset (MIND)

- 1.000.000 de usuarios (junto a sus interacciones).
- Más de 160.000 noticias.
- Datos recogidos a lo largo de 6 semanas.

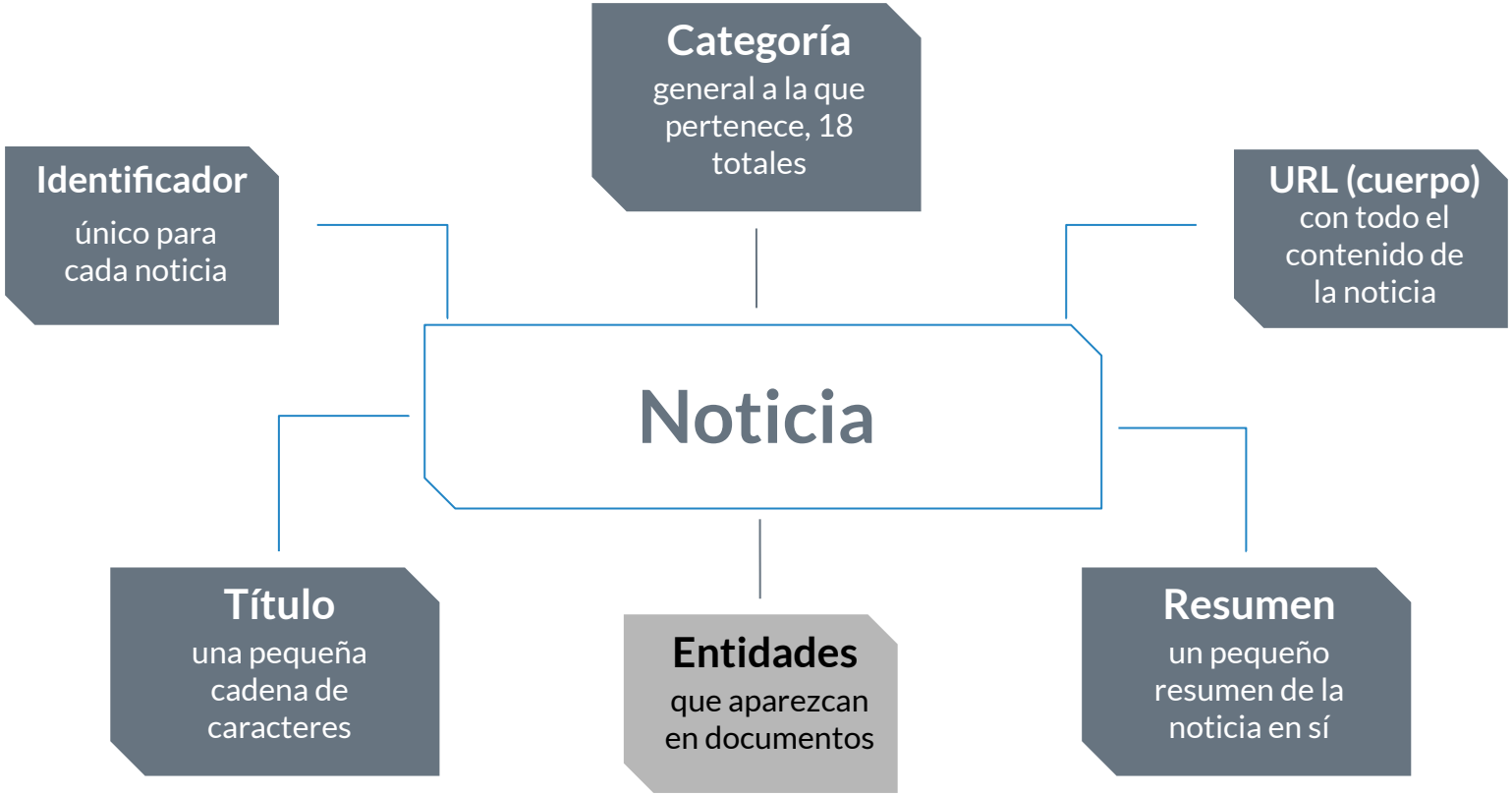
Nº Muestras	**train_news**	**valid_news**	**train_behaviors**	**valid_behaviors**
demo	26.740	18.723	21.480	7.335
small	51.282	42.416	153.727	70.938
large	101.527	72.023	2.186.683	365.201

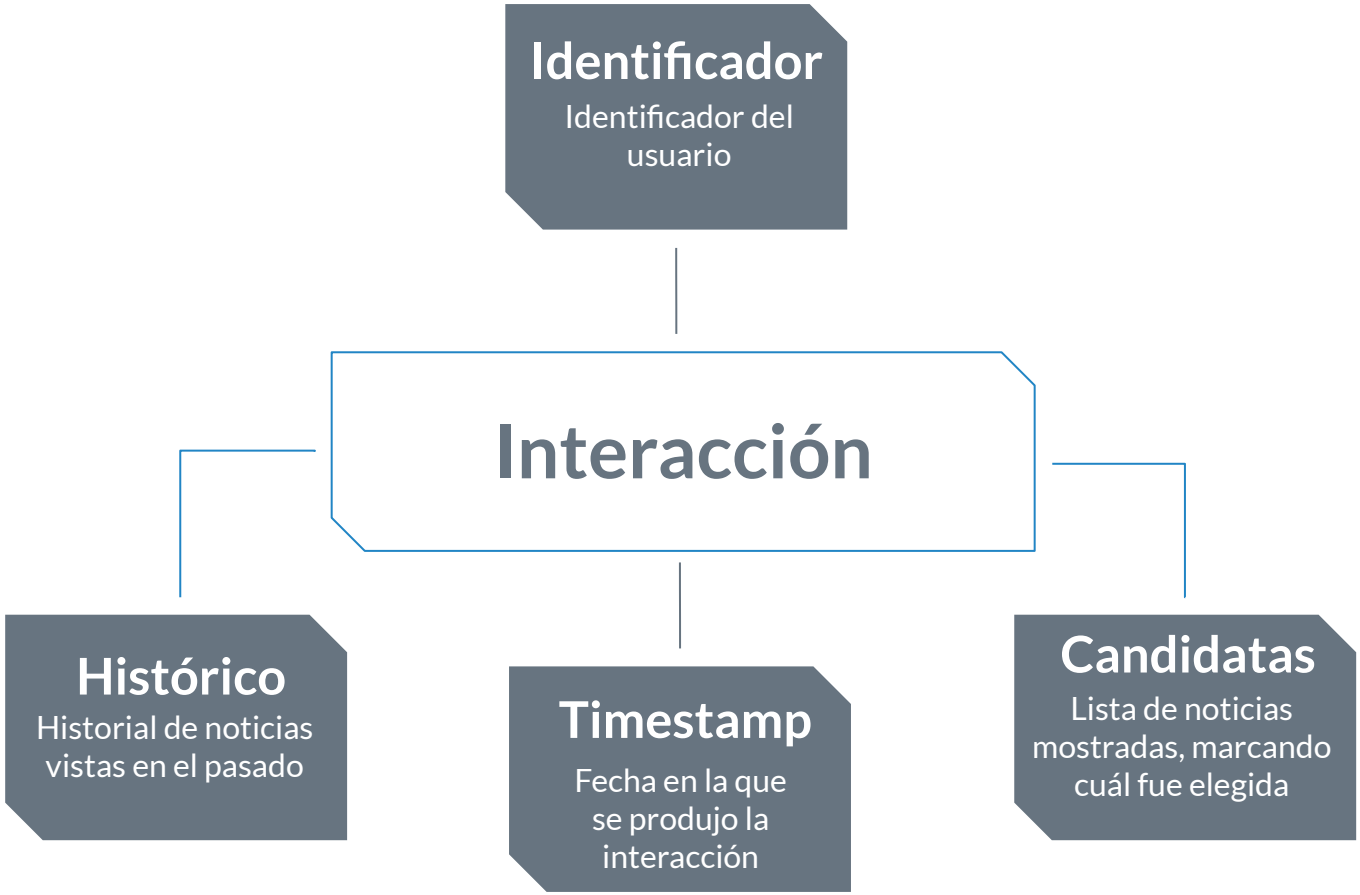
MIcrosoft News Dataset (MIND)

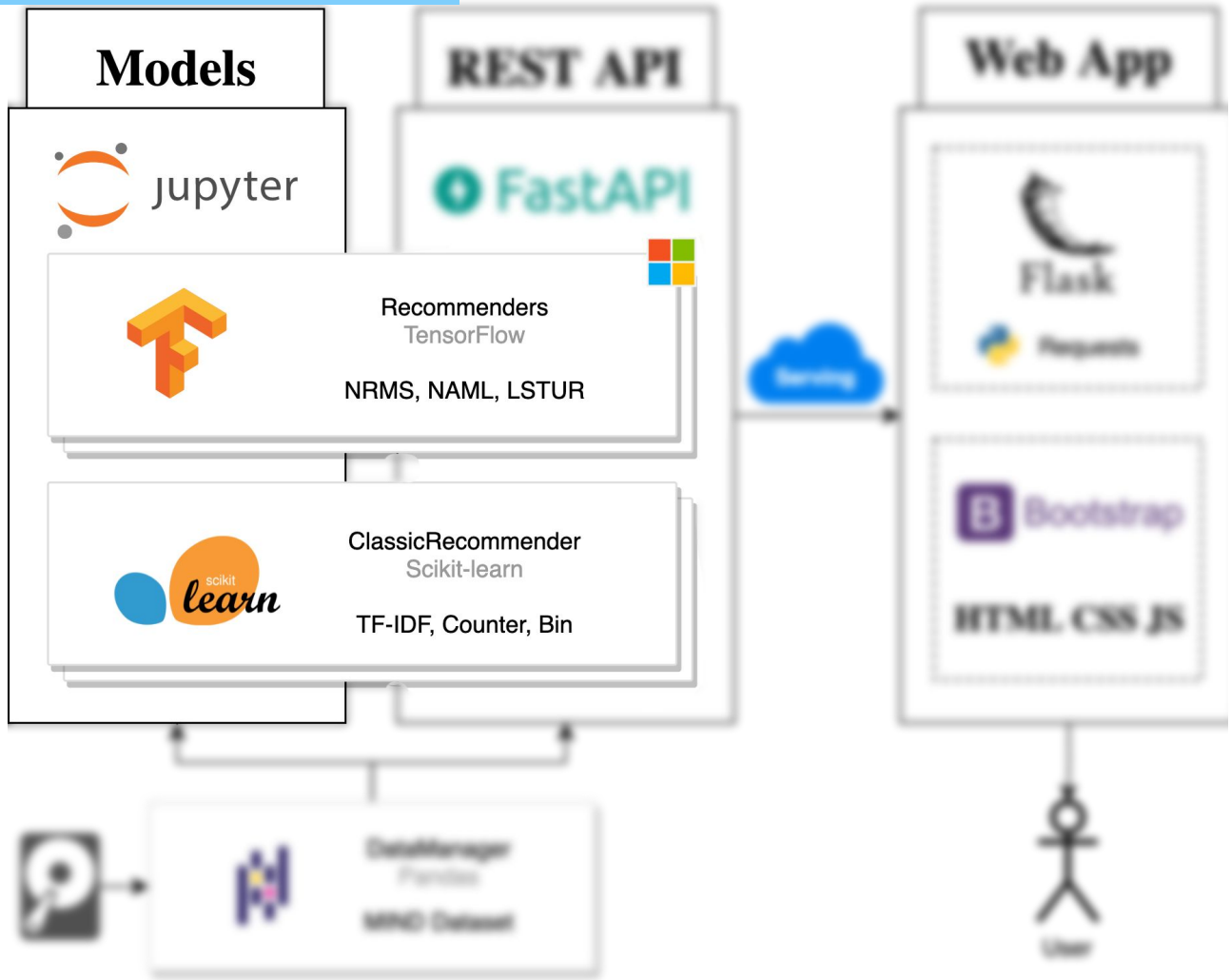
- 1.000.000 de usuarios (junto a sus interacciones).
- Más de 160.000 noticias.
- Datos recogidos a lo largo de 6 semanas.

~4.000 noticias diarias

Inabarcable para cualquier persona


Categoría
general a la que pertenece, 18 totales
Identificador
único para cada noticia
URL (cuerpo)
con todo el contenido de la noticia
Noticia
Título
una pequeña cadena de caracteres
Entidades
que aparezcan en documentos
Resumen
un pequeño resumen de la noticia en sí


Identificador
Identificador del usuario
Interacción
Histórico
Historial de noticias vistas en el pasado
Timestamp
Fecha en la que se produjo la interacción
Candidatas
Lista de noticias mostradas, marcando cuál fue elegida


Models
REST API
Web App
jupyter
FastAPI
Flask
Recommenders
TensorFlow
NRMS, NAML, LSTUR
Bootstrap
HTML CSS JS
ClassicRecommender
Scikit-learn
TF-IDF, Counter, Bin

Modelos

ClassicRecommender

Modelos basados en algoritmos clásicos, TF-IDF, frecuencia de palabras...

Recommenders

Modelos de Aprendizaje Profundo. Adaptados del repositorio *recommenders* de Microsoft.

https://github.com/microsoft/recommenders

ClassicRecommender

Scikit-learn

TF-IDF, Counter, Bin

ClassicRecommender

Vectorización

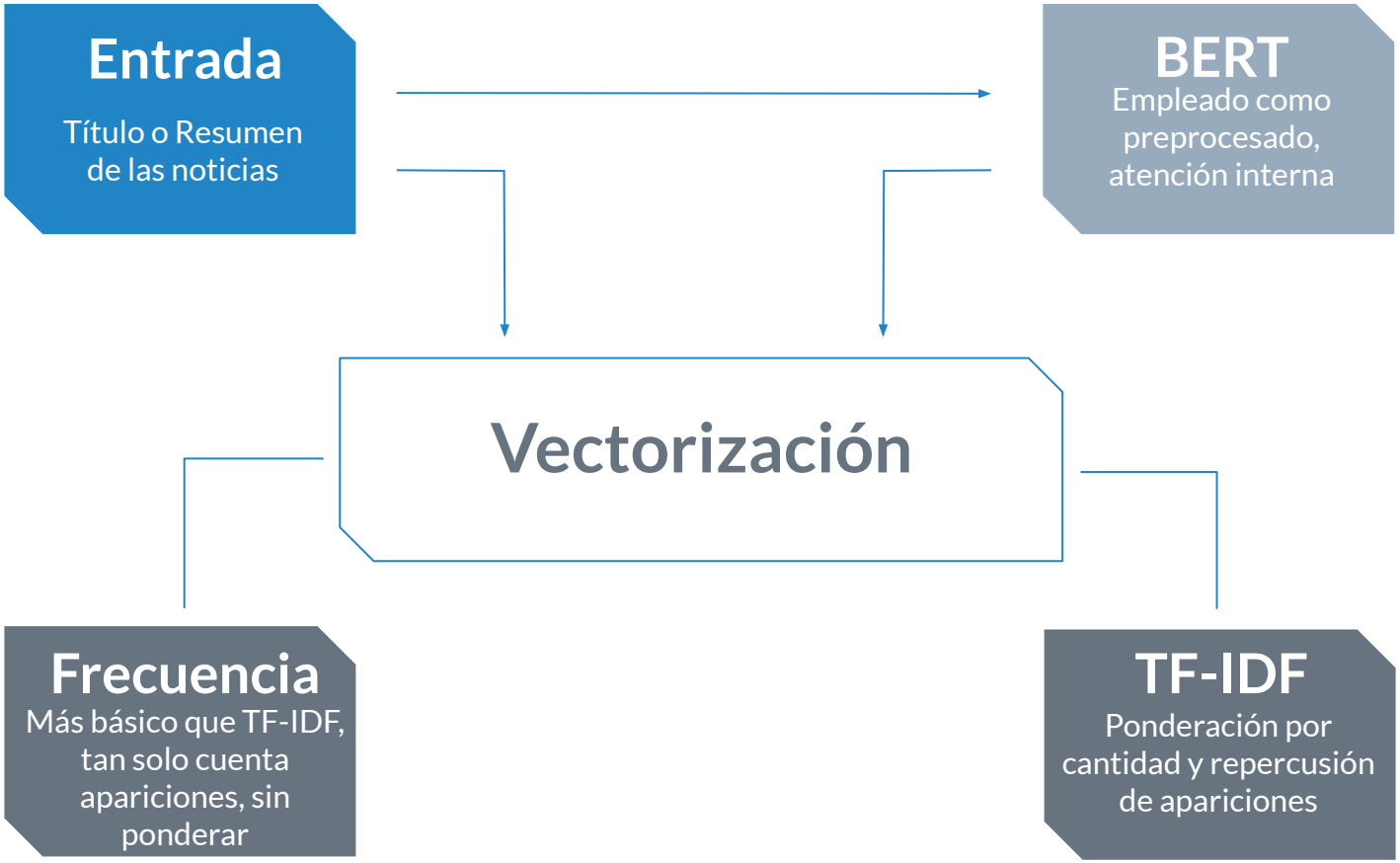
Construir una representación numérica a partir de un texto.

Bienvenido	al	trabajo	de	Eric	Morales
2	1	7	3	4	6

Similitud

Se calculan las similitudes entre las noticias visitadas en el pasado y las candidatas.

Definiendo la probabilidad en función de la similitud.


Entrada
Título o Resumen de las noticias
BERT
Empleado como preprocesado, atención interna
Vectorización
Frecuencia
Más básico que TF-IDF, tan solo cuenta apariciones, sin ponderar
TF-IDF
Ponderación por cantidad y repercusión de apariciones

Recommenders

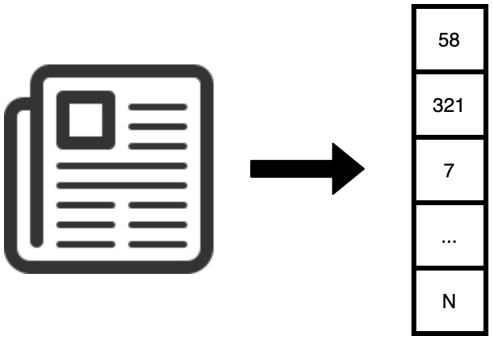
TensorFlow

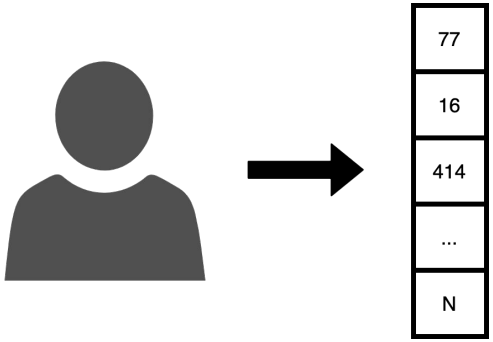
NRMS, NAML, LSTUR

Recommenders

Encoder de noticias

Utilizando su contexto, contenido, categoría, etc.





Encoder de usuarios

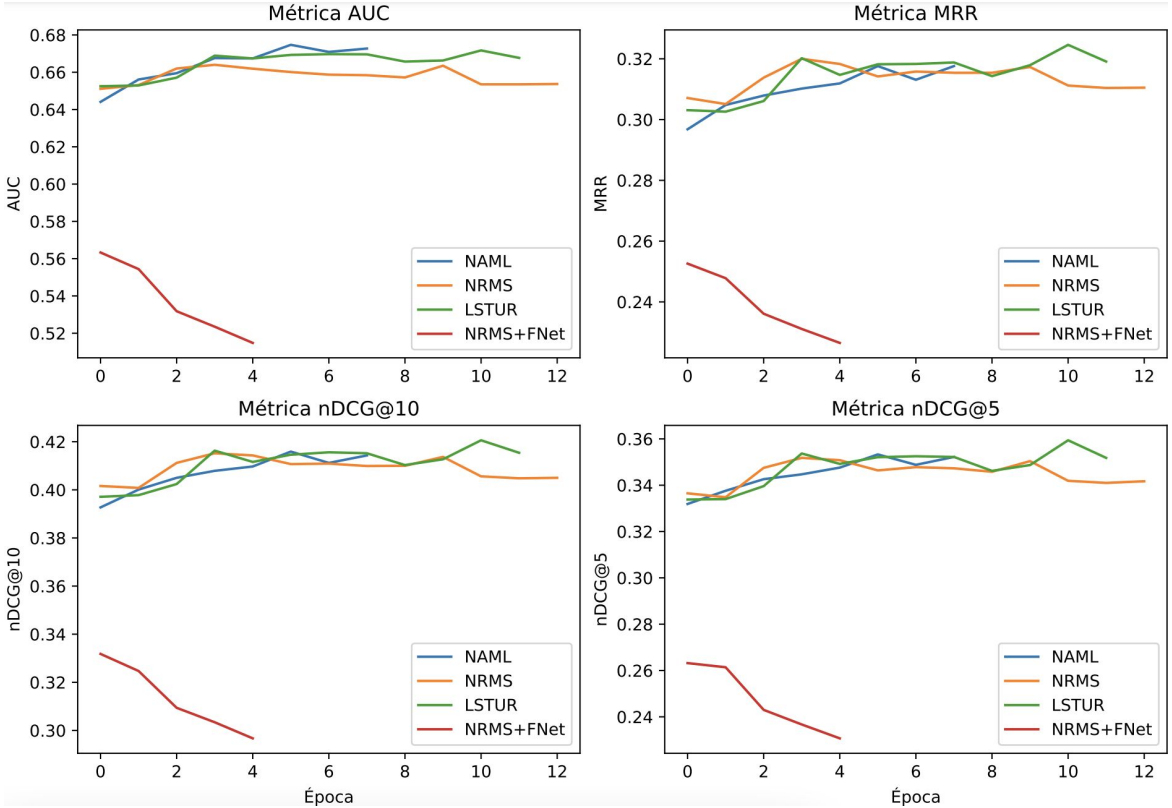
En base a otros usuarios y/o a las noticias que este ha visto.

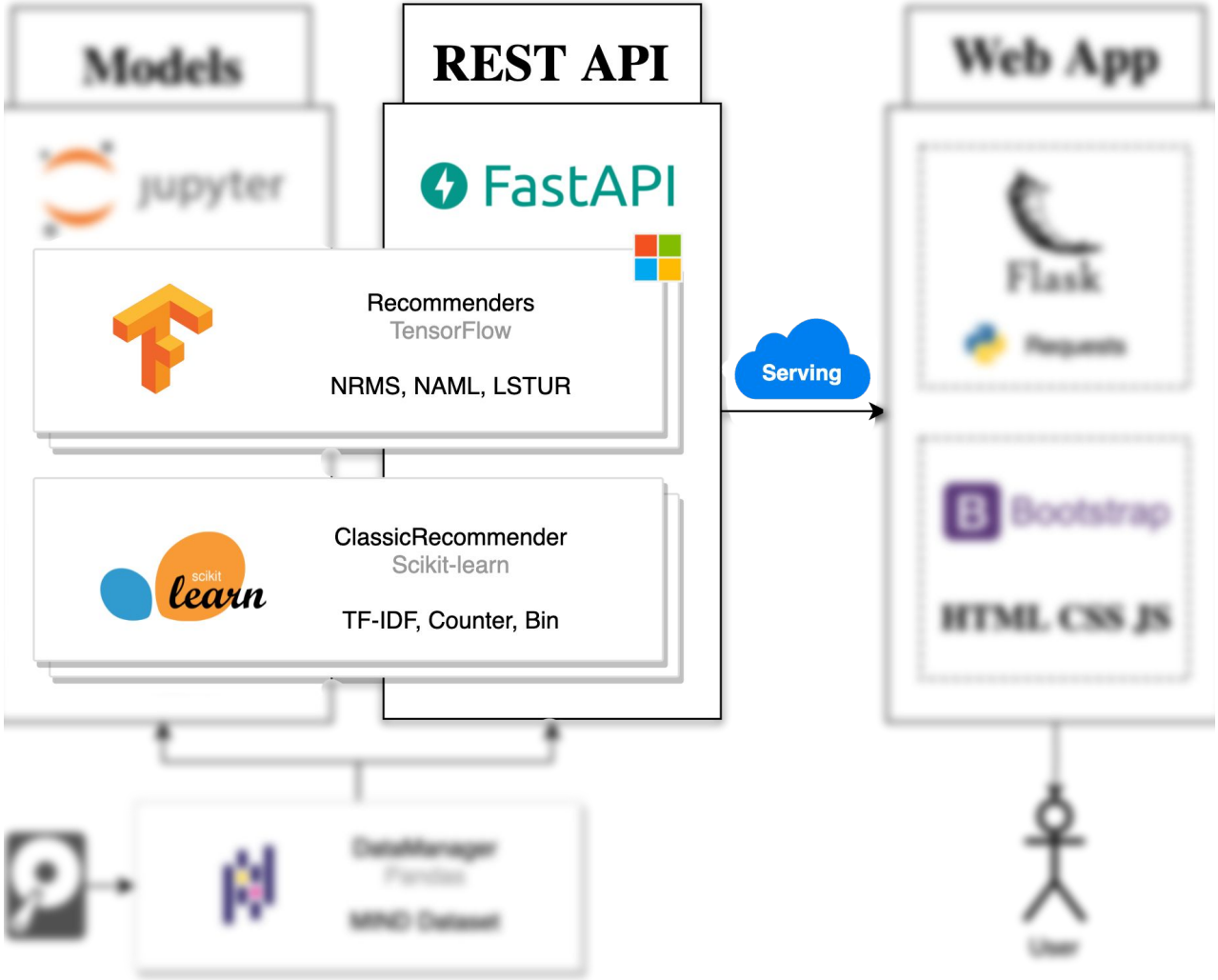
Recommenders

	NRMS	NAML	LSTUR
NOTICIAS	Título. **Multi-head self-attention** para capturar relaciones entre palabras.	Título, cuerpo y categoría. Representación multi-view, **atención** para las palabras y views más importantes.	Título y categoría. Aplica **atención** para capturar las palabras más importantes.
USUARIOS	**Multi-head self-attention** para capturar las relaciones entre noticias visitadas.	**Atención** para seleccionar las noticias visitadas más importantes.	Dos encoders, uno a corto plazo (**GRU**) y otro a largo plazo (**embeddings**).

Entrenamiento

Resultados Recommender



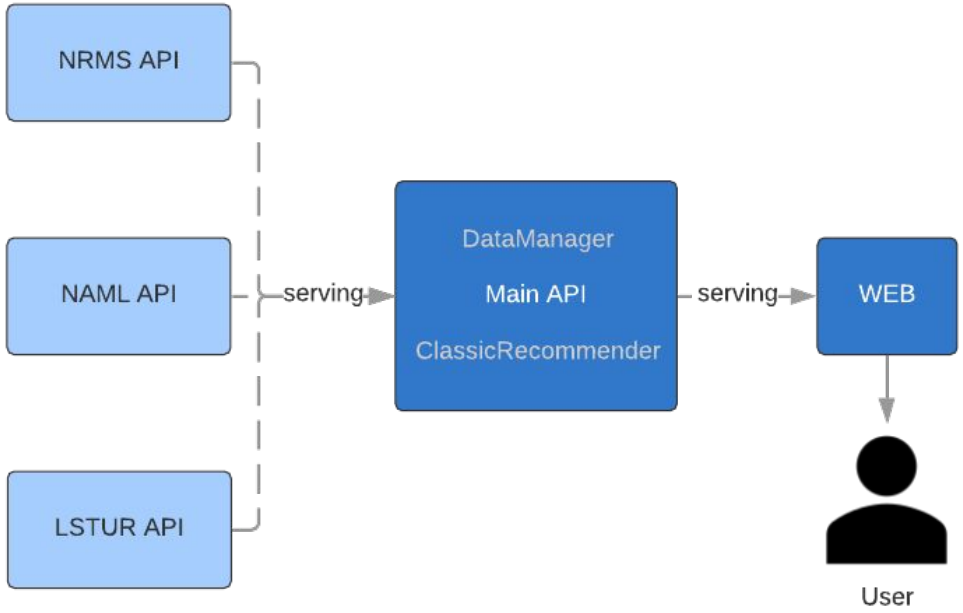

Models
REST API
Web App
jupyter
FastAPI
Recommenders
TensorFlow
NRMS, NAML, LSTUR
Serving
Flask
Bootstrap
HTML CSS JS
ClassicRecommender
Scikit-learn
TF-IDF, Counter, Bin

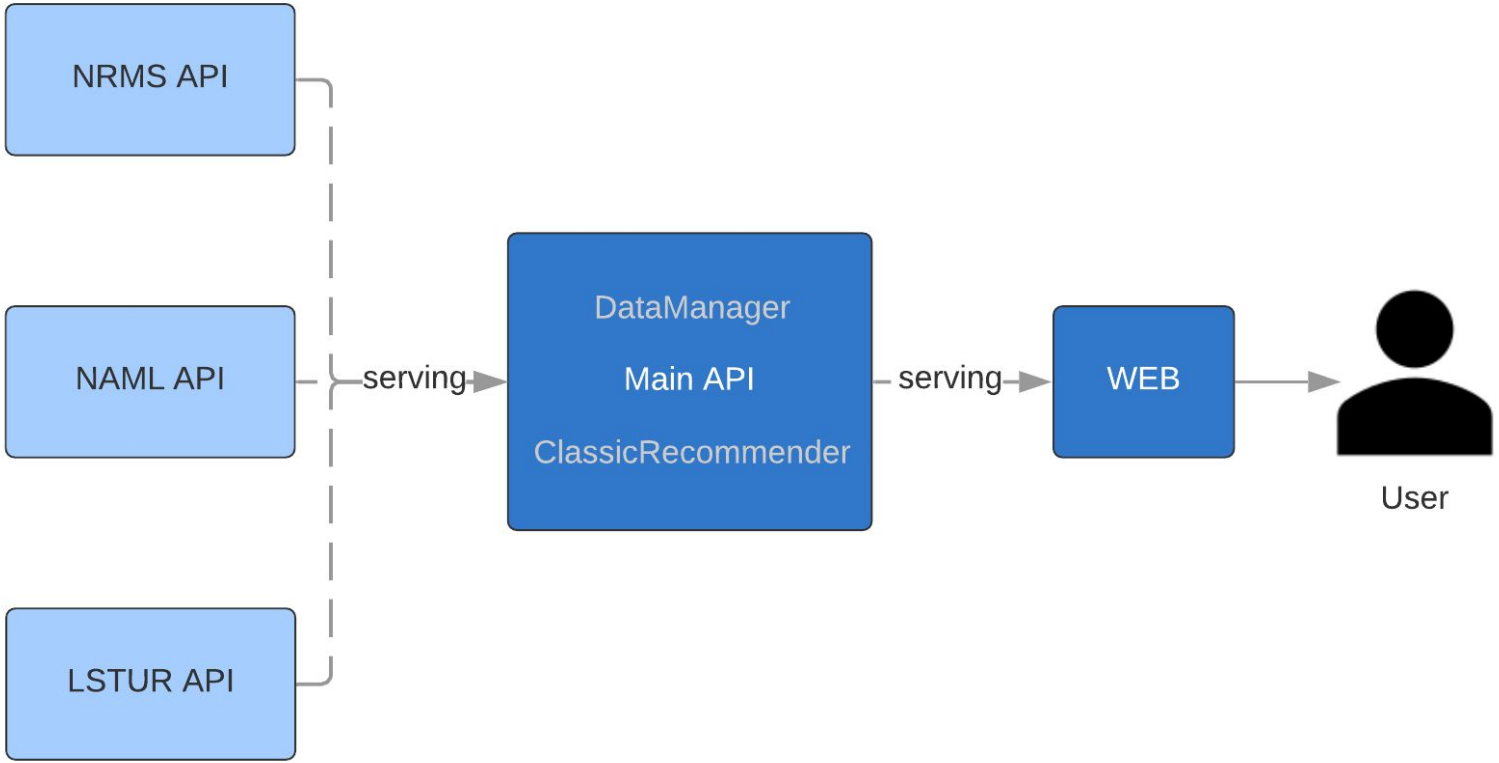
REST API

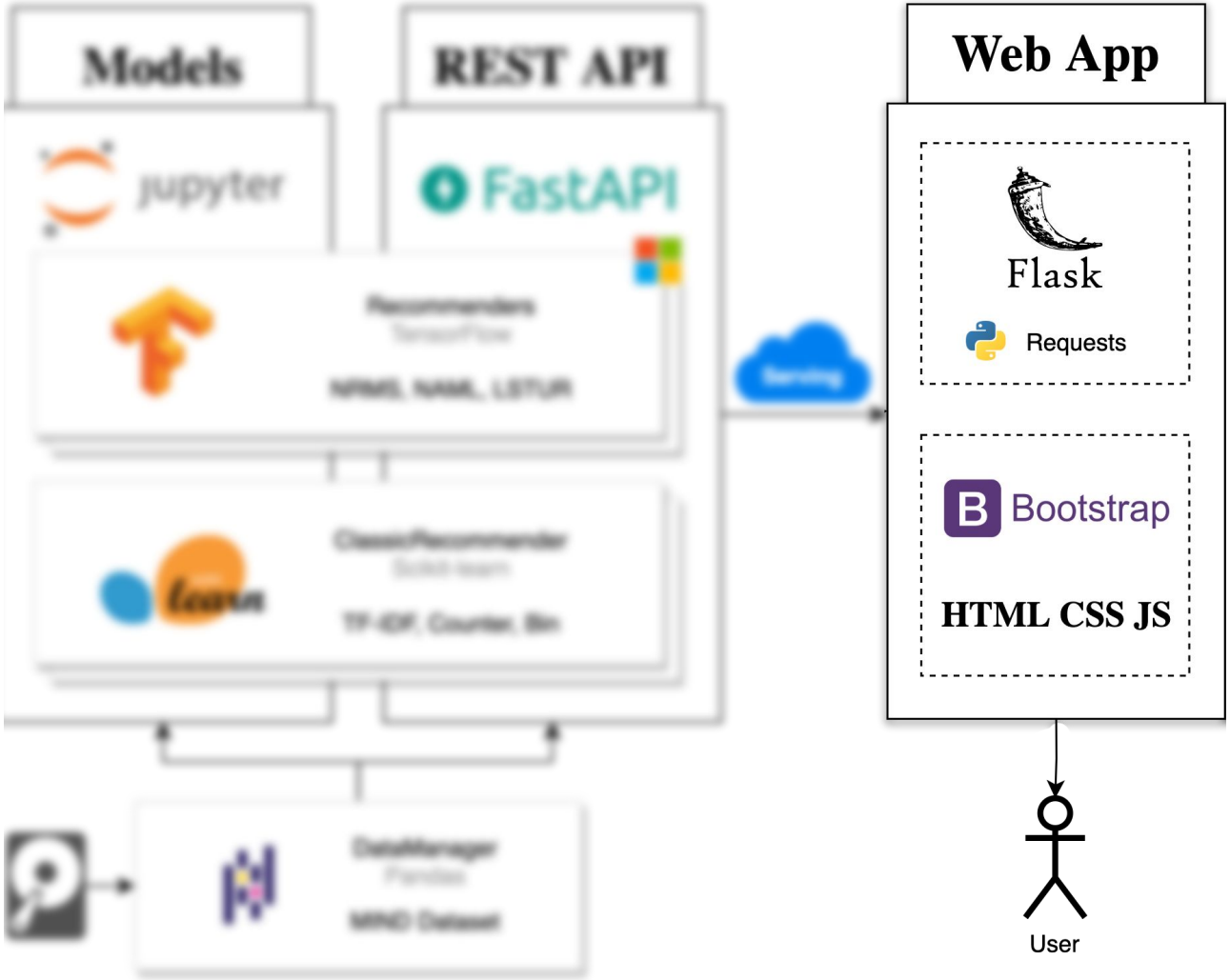

GET /recommendation/ Get Recommendation
GET /available_news/ Get Available News
GET /clicked_news/ Get Clicked News
GET /timestamp/ Get Timestamps
GET /users_sample/ Get Users Sample
GET /news_sample/ Get News Sample

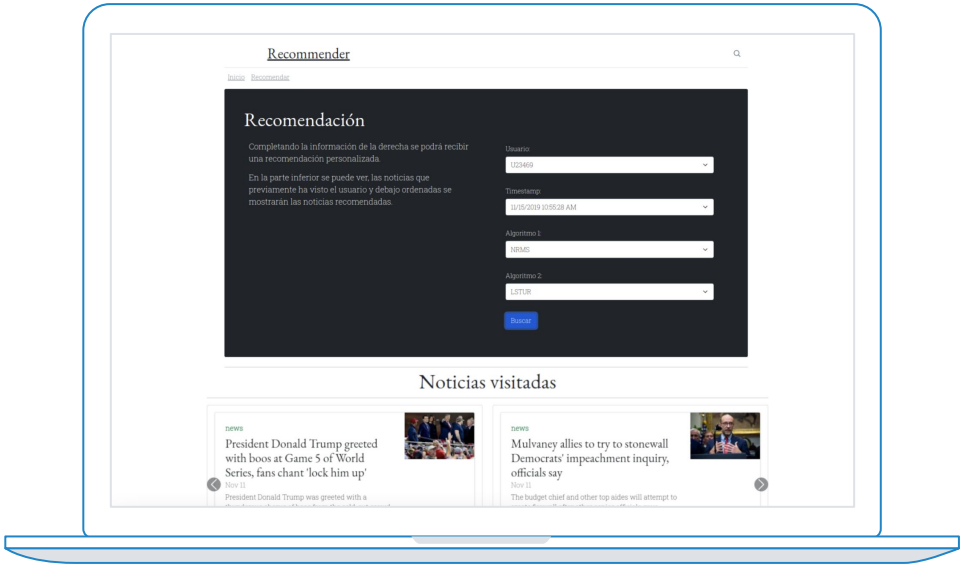
Cuatro API REST

La realidad es que, debido a problemas internos de Tensorflow y CUDA, fue necesario hacer una API para cada modelo (3), además de la principal.




NRMS API
NAML API
LSTUR API
serving
DataManager
Main API
ClassicRecommender
serving
WEB
User


Models
REST API
Web App
jupyter
FastAPI
Recommenders
TensorFlow
NRMS, NAML, LSTUR
Serving
Flask
Requests
Bootstrap
HTML CSS JS
ClassicRecommender
Scikit-learn
TF-IDF, Counter, Bin
DataManager
Pandas
MIND Dataset
User



Página Web

Se trata de una web que consume de la API mencionada, utilizando requests Python y AJAX

Recommender

Inicio Recomendar

Recomendación

Completando la información de la derecha se podrá recibir una recomendación personalizada.

En la parte inferior se puede ver, las noticias que previamente ha visto el usuario y debajo ordenadas se mostrarán las noticias recomendadas.

Usuario:

U23469

Timestamp:

11/15/2019 10:55:28 AM

Algoritmo 1:

NRMS

Algoritmo 2:

LSTUR

Buscar

Noticias visitadas

news

President Donald Trump greeted with boos at Game 5 of World Series, fans chant 'lock him up'

Nov 11

President Donald Trump was greeted with a

news

Mulvaney allies to try to stonewall Democrats' impeachment inquiry, officials say

Nov 11

The budget chief and other top aides will attempt to

3. CONCLUSIONES

Conclusiones y trabajo futuro

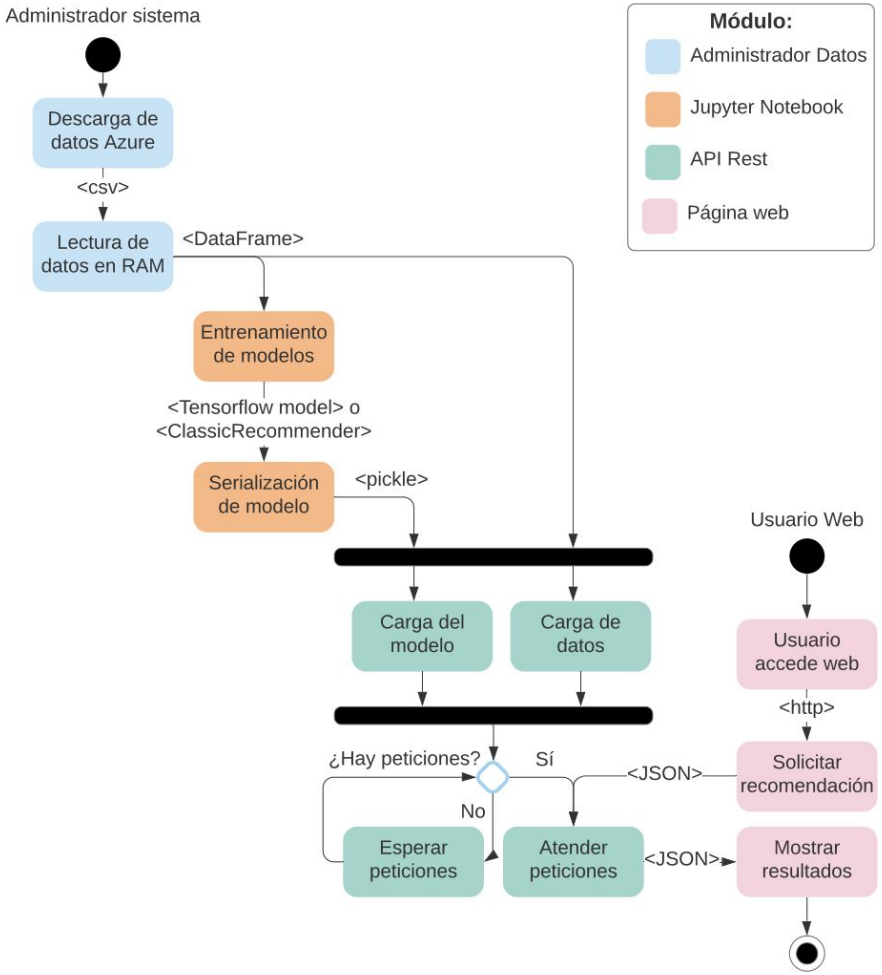
Conclusiones

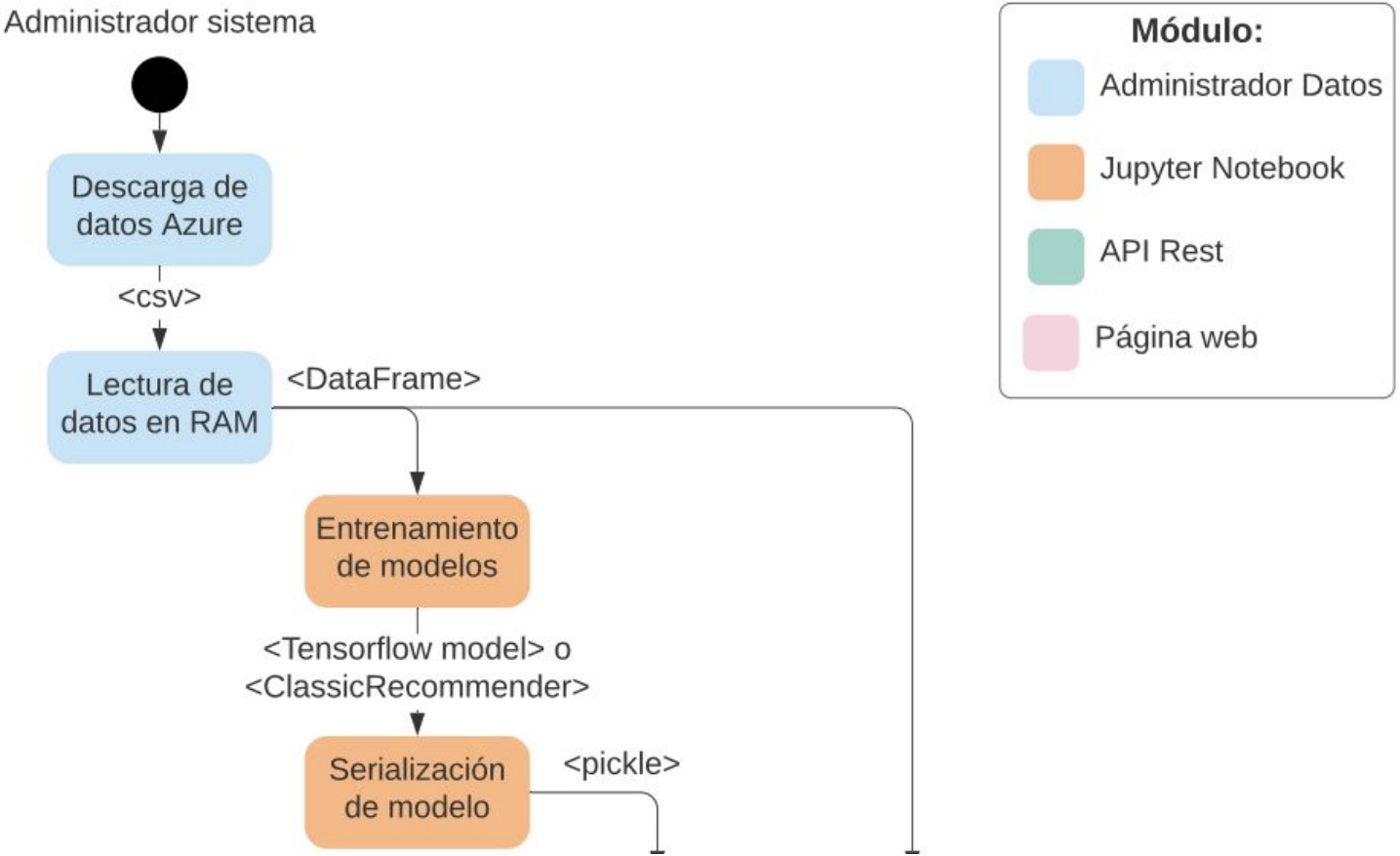
Desarrollo Aplicación

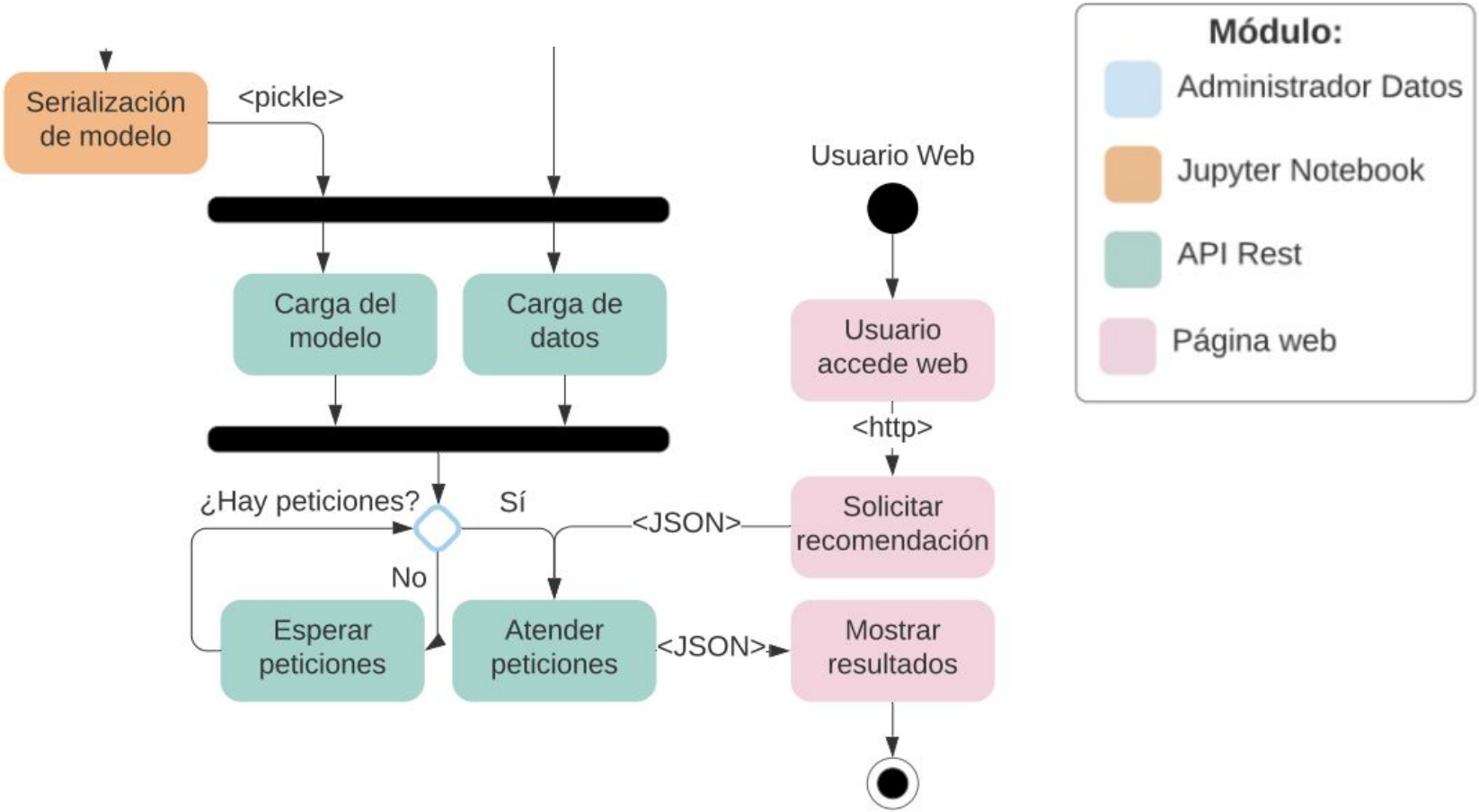
Infraestructura completa, desde la lectura de los datos hasta presentación en la página web.

Investigación Algoritmos

Desarrollo de algoritmos propios, adaptación de algoritmos externos y reproducción de experimentos.


Módulo:
Administrador Datos
Jupyter Notebook
API Rest
Página web
Administrador sistema
Descarga de datos Azure
<csv>
Lectura de datos en RAM
<DataFrame>
Entrenamiento de modelos
<Tensorflow model> o <ClassicRecommender>
Serialización de modelo
<pickle>
Carga del modelo
Carga de datos
¿Hay peticiones?
Sí
No
Esperar peticiones
Atender peticiones
Usuario Web
Usuario accede web
<http>
Solicitar recomendación
<JSON>
Mostrar resultados
<JSON>


Administrador sistema
Descarga de datos Azure
<csv>
Lectura de datos en RAM
<DataFrame>
Entrenamiento de modelos
<Tensorflow model> o
<ClassicRecommender>
Serialización de modelo
<pickle>
Módulo:
Administrador Datos
Jupyter Notebook
API Rest
Página web


Módulo:
Administrador Datos
Jupyter Notebook
API Rest
Página web
Serialización de modelo
<pickle>
Carga del modelo
Carga de datos
Usuario Web
Usuario accede web
<http>
Solicitar recomendación
<JSON>
¿Hay peticiones?
Sí
No
Esperar peticiones
Atender peticiones
<JSON>
Mostrar resultados

Trabajo futuro: Modelo

Optimizar FNet

Optimizar la implementación del modelo NRMS que en lugar de utilizar atención utiliza transformadas de Fourier discretas (FFT).

Uso de imágenes

Incorporar imágenes a alguno de los algoritmos de recomendación descritos, probablemente NAML, utilizando redes CNN.

Muchas gracias

Turno de preguntas

Eric Morales Agostinho
Ingeniería Informática
Escuela Politécnica Superior

ANEXO A.
CLASSIC RECOMMENDER

Detalles técnicos de la implementación

ClassicRecommender

Vocabulario (al azar):	
a	0
al	1
bienvenido	2
de	3
Eric	4
horario	5
Morales	6
trabajo	7
vaca	8
ukelele	9
...	N

Bienvenido al trabajo de vacas de Eric Morales

0	1	1	2	1	0	1	1	1	0	...
0	0.01	1	0.001	1	0	1	1	1	0	...

Frecuencia

TF-IDF

No es el resultado real

$$tf(t,d) = \begin{cases} 1 + log_2\ freq(t,d), & \text{si } freq(t,d) > 0 \\ 0, & \text{en otro caso} \end{cases}$$

$$idf(t) = log\frac{|\mathcal{D}|}{\mathcal{D}_t}$$

$$w_{tf-idf}(i,j) = tf(t_i,d_j) \times idf(t_i)$$

ANEXO B.
DETALLES RECOMMENDERS

Ejemplos y motivación de algoritmos recommenders

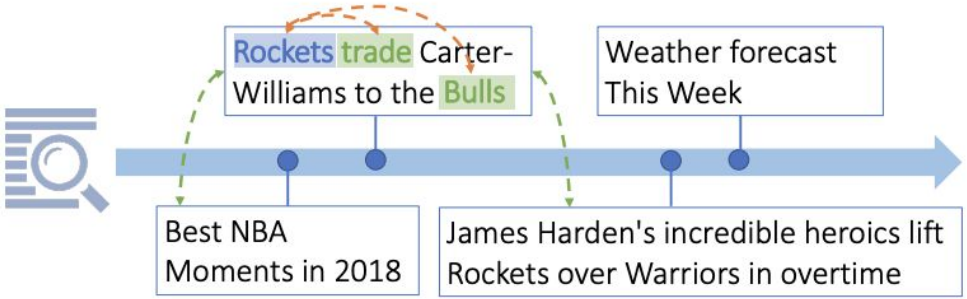
NRMS [1]

Encoder de noticias

Utilizando el título, aplica multi-head self-attention para capturar relaciones entre palabras.

Encoder de usuarios

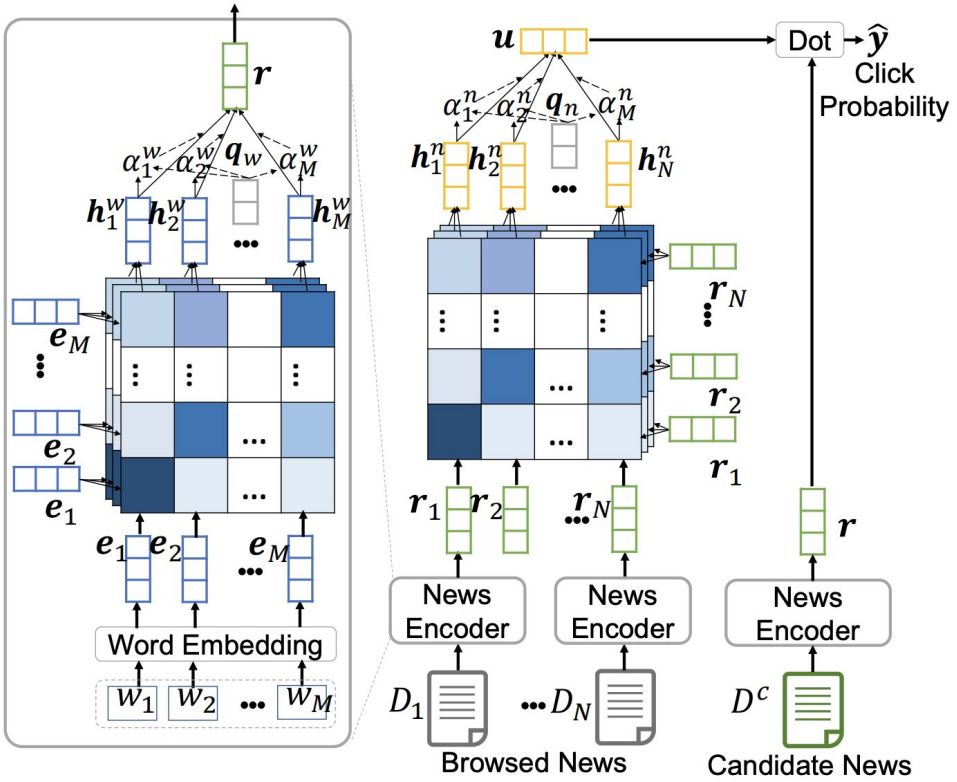
En base a las noticias visitadas en el pasado, utiliza multi-head self-attention para capturar las relaciones entre ellas.



[1] Wu, Chuhan, et al. "Neural news recommendation with multi-head self-attention." in EMNLP-IJCNLP. 2019.

NRMS

Neural News Recommendation with Multi-Head Self-Attention



[1] Wu, Chuhan, et al. "Neural news recommendation with multi-head self-attention." in EMNLP-IJCNLP. 2019.

NAML [2]

Encoder de noticias

Utilizando el título, cuerpo y categoría, genera una representación multidimensional, utiliza atención para las palabras y views más importantes.

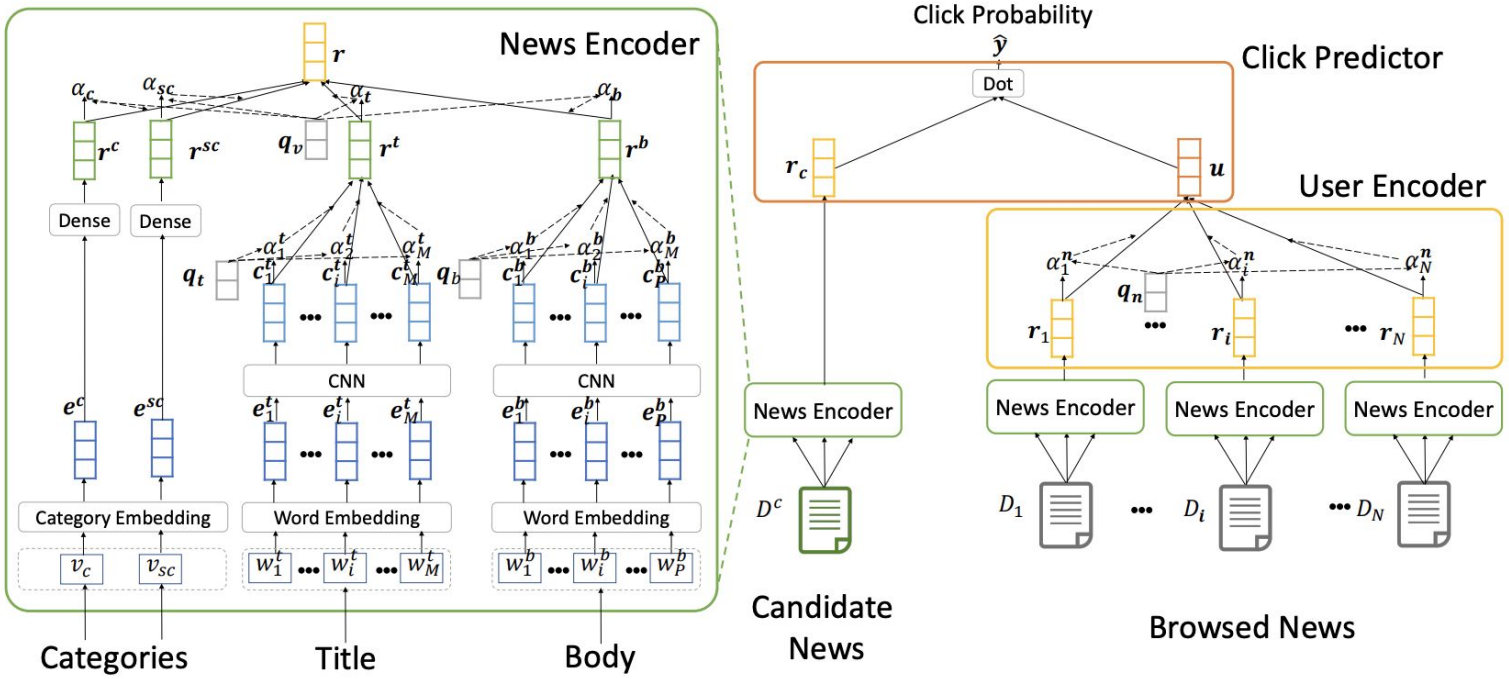
Encoder de usuarios

En base a las noticias visitadas en el pasado, utiliza atención para seleccionar las más importantes.

Category	Sports	Entertainment
Title	Astros improve outfield, agree to 2-year deal with Brantley	The best games of 2018
Body	Outfielder Michael Brantley agreed to a two-year, $32 million contract with Houston, sources familiar with the deal told Yahoo Sports, bringing his steady left-handed bat to the top of an Astros ...	The Best Games of 2018 Superheroes, super-dads, and Super Mario parties brought the joy in 2018 to millions of players who ate up the year's impressive achievements in gaming ...

[2] C. Wu, et al. “Neural news recommendation with attentive multi-view learning,” in IJCAI, 2019.

NAML: Neural News Recommendation with Attentive Multi-View Learning



[2] C. Wu, et al. "Neural news recommendation with attentive multi-view learning," in IJCAI, 2019.

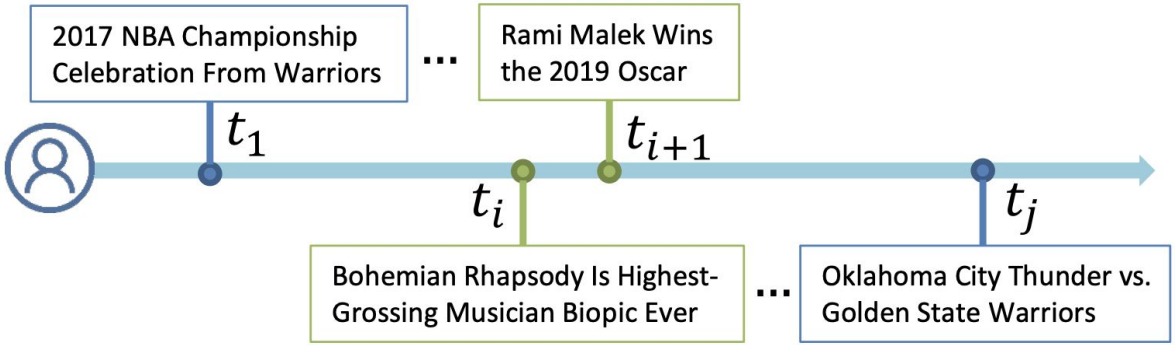
LSTUR [3]

Encoder de noticias

Utilizando el título y categoría,, aplica atención para capturar las palabras más importantes.

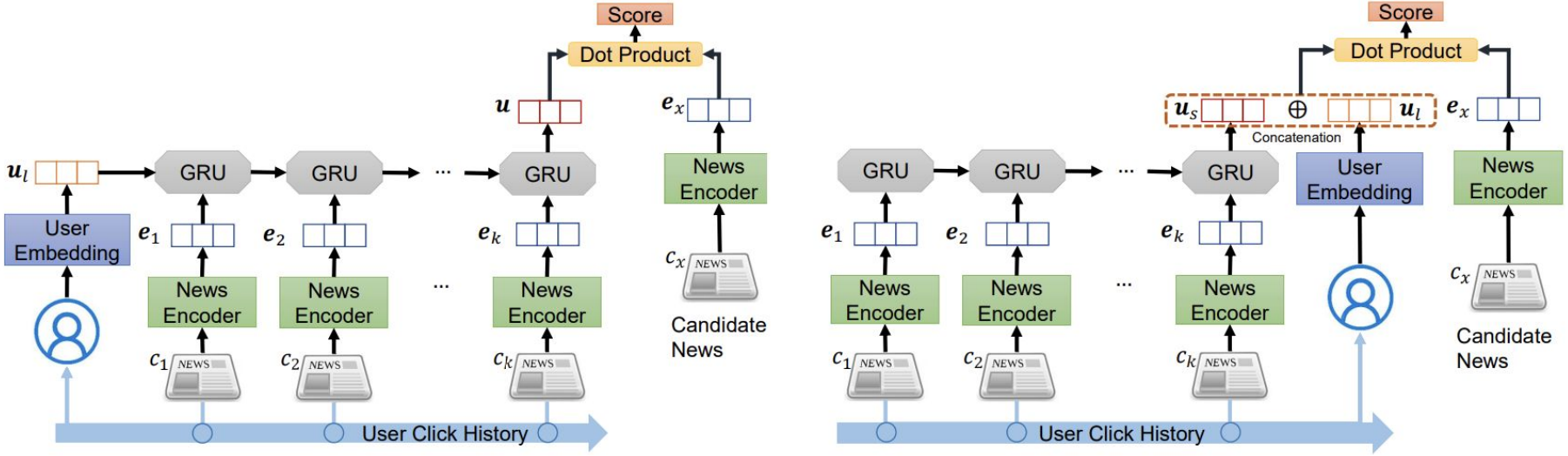
Encoder de usuarios

Se divide en dos encoder, uno a corto plazo y otro a largo plazo. Utilizando redes GRU a corto plazo y embeddings de IDs a largo plazo.



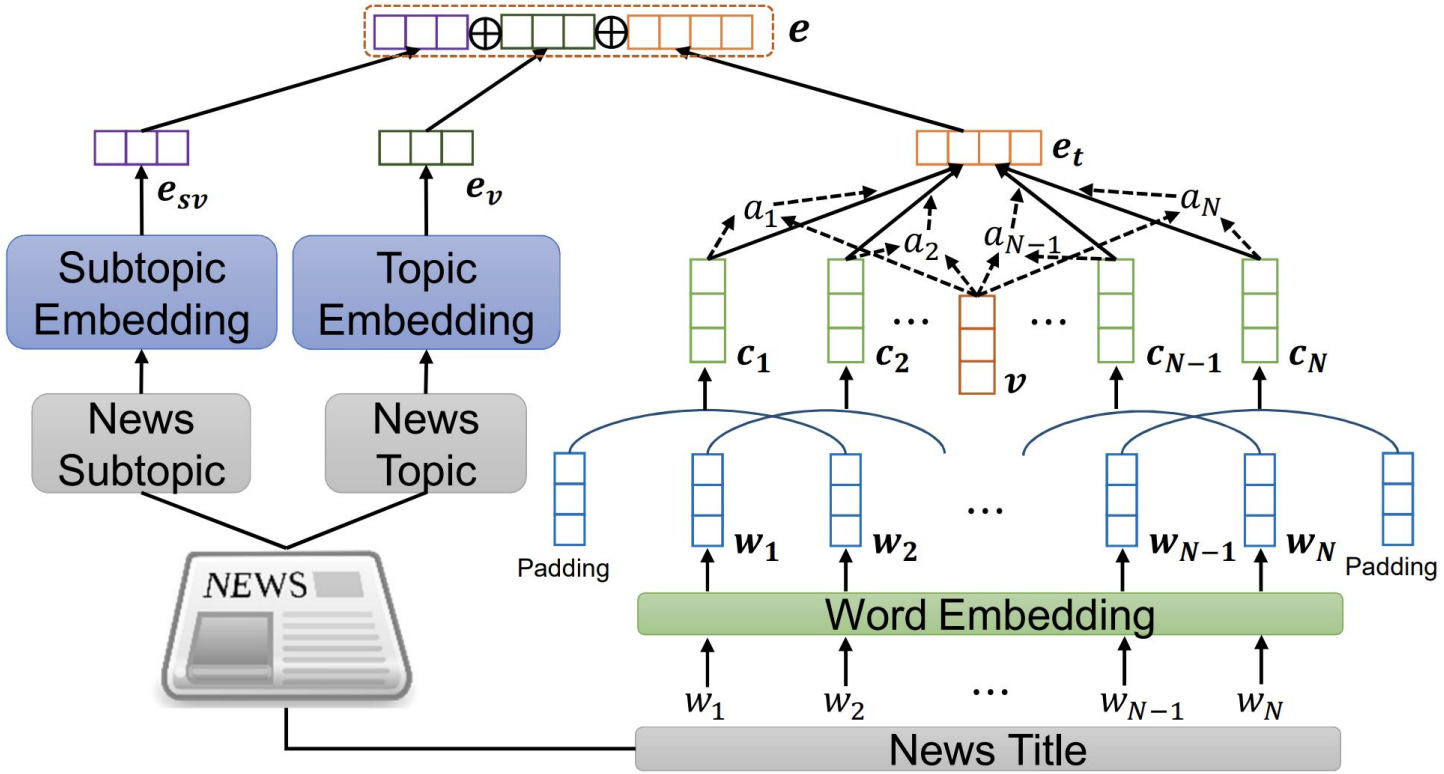
[3] M. An, et al. “Neural news recommendation with long- and short-term user representations,” in ACL, 2019.

LSTUR: Long- and Short-term User Representations



[3] M. An, et al. “Neural news recommendation with long- and short-term user representations,” in ACL, 2019.

LSTUR: Long- and Short-term User Representations



[3] M. An, et al. "Neural news recommendation with long- and short-term user representations," in ACL, 2019.

ANEXO C.
TABLAS RESULTADOS

Tablas con resultados detallados

Resultados ClassicRecommender

Algoritmo	*Sin BERT (Dataset completo)*			*Con BERT (500 muestras)*		
Frecuencia	**Sim. coseno**	**Euclídea**	**Manhattan**	**Sim. coseno**	**Euclídea**	**Manhattan**
AUC	**0,5015**	0,4977	0,4976	0,4856	0,4861	0,4902
MRR	0,2190	0,2173	0,2173	0,2423	0,2756	**0,2773**
nDCG@5	0,2237	0,2222	0,2220	0,2259	0,2923	**0,2995**
nDCG@10	0,2866	0,2850	0,2852	**0,3325**	0,3304	0,3315

Algoritmo	*Sin BERT (Dataset completo)*			*Con BERT (500 muestras)*		
TF-IDF	**Sim. coseno**	**Euclídea**	**Manhattan**	**Sim. coseno**	**Euclídea**	**Manhattan**
AUC	**0,5005**	0,5004	0,4977	0,4743	0,4776	0,4997
MRR	0,2191	0,2191	0,2175	0,2606	0,2619	**0,2729**
nDCG@5	0,2238	0,2238	0,2223	0,2493	0,2647	**0,2997**
nDCG@10	0,2865	0,2865	0,2853	0,3397	**0,3407**	0,3392

Resultados Recommender

Algoritmo	Época	AUC	MRR	nDCG@10	nDCG@5
NAML	6	**0,6747**	**0,3176**	0,4159	0,3533
NRMS	4	0,6640	0,3200	0,4152	0,3518
LSTUR	11	0,6717	0,3246	**0,4206**	**0,3594**
NRMS+FNet	1	0,5633	0,2526	0,3318	0,2632