
quantil

Algunas aplicaciones de aprendizaje de máquinas en Colombia

Carlos Andrés Reyes
Quantil – Universidad del Valle

Mieroles 17 de Octubre del 2018

Una nueva cultura

- En los últimos 2 años se ha producido y almacenado más información que en toda la historia de la humanidad.
- El acelarador de partículas en CERN puede generar varios petabytes (10^6 GB) de información diaria.
- Con los datos recolectados en las redes sociales Facebook nos conoce mejor que nuestros familiares.

Colombia no es la excepción

- Los negocios buscan hacerle seguimiento a sus clientes e implementar politicas para retenerlos o hacerles ofertas.
- Las instuticiones financieras desean tener un cálculo más preciso de los riesgos a los que están expuestas.
- Los grupos de investigación tienen acceso a nuevas bases de datos para expandir sus proyectos.

Gapmaps

- Aproximadamente 60,000 artículos para cada tema.
- ¿Podemos clasificar los artículos en un número dado de tópicos?
- ¿Cuales son los tópicos más interesantes?
- ¿Que metodologías se utilizan en los artículos?

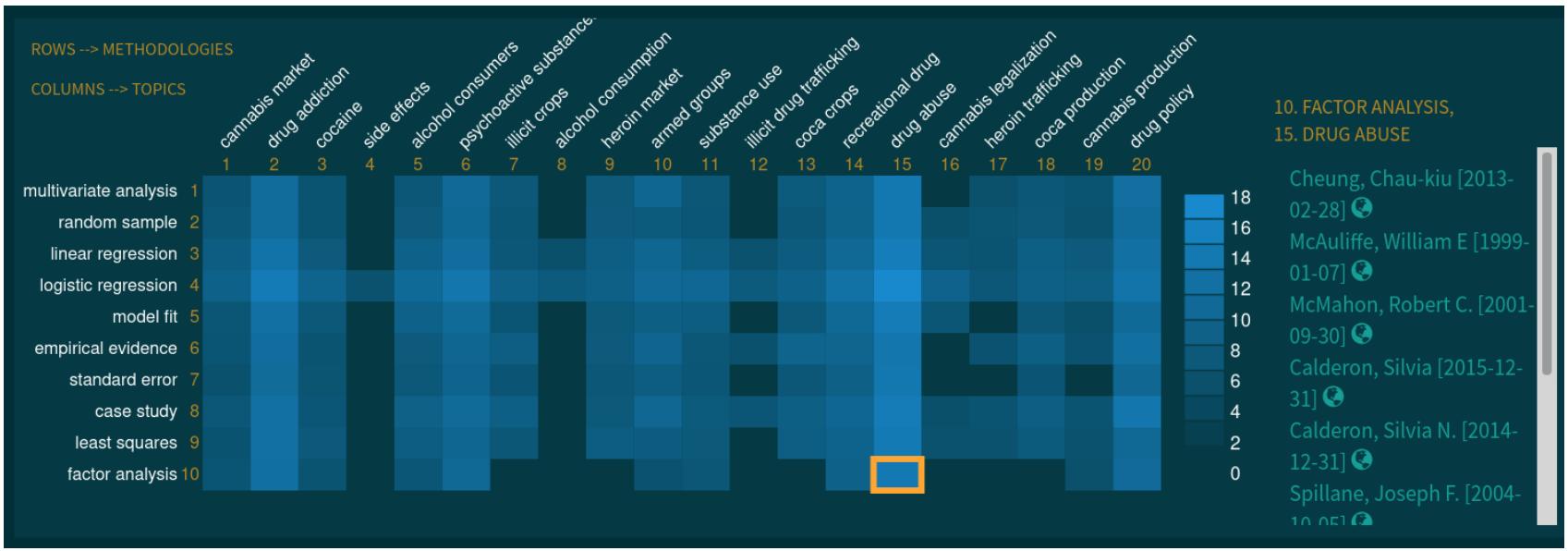


Figura: Gapmap para 60,000 artículos sobre drogas.

Riesgo de Crédito

- Se desea calcular la probabilidad de que alguien incumpla con los pagos de un crédito.
- Se conocen algunos datos del cliente al momento de pedir el crédito (variables demográficas, comportamiento crediticio).
- Se hace un seguimiento del comportamiento de pago del cliente para así calcular su probabilidad de incumplimiento en cualquier momento.

Redes neuronales recurrentes para las series de tiempo

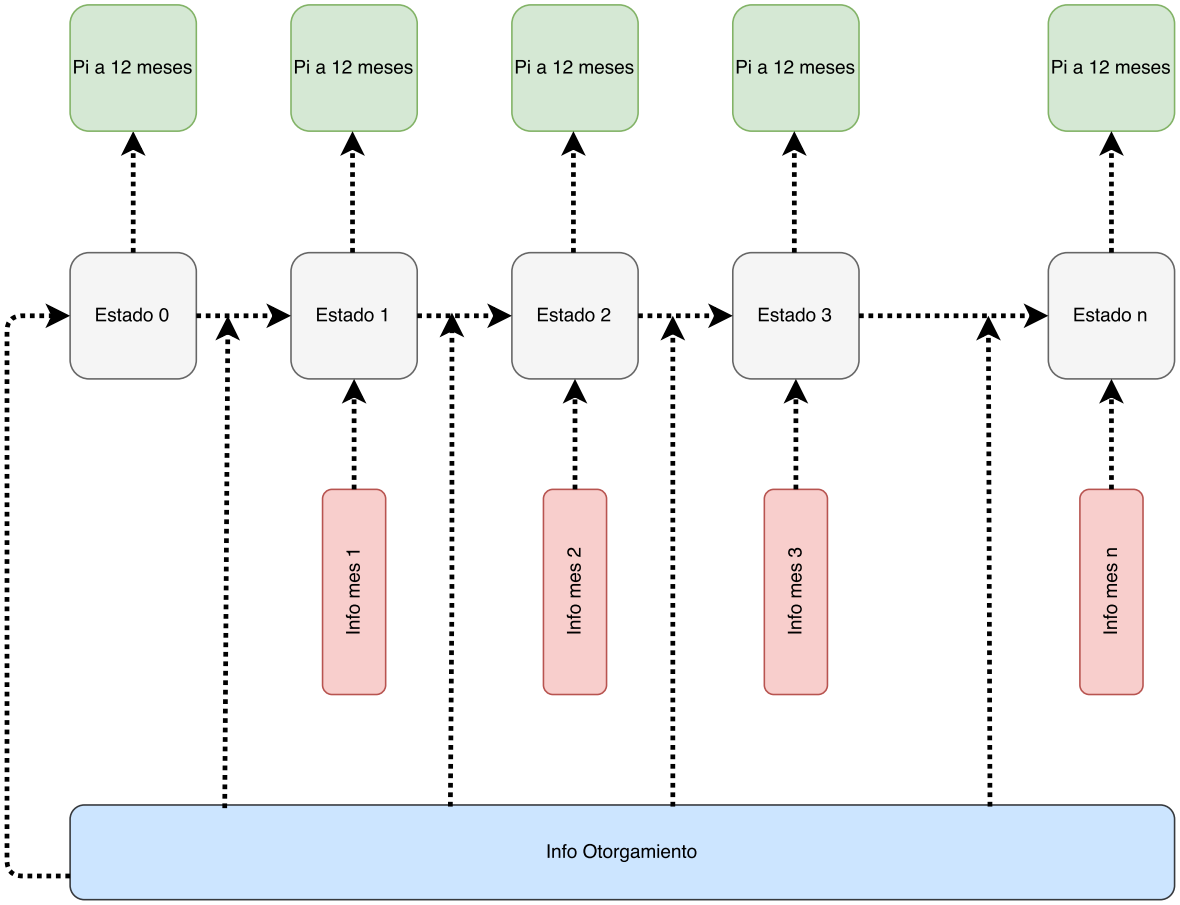


Figura: Diagrama de una red neuronal recurrente en el ámbito de riesgo de crédito.

Neuronas LSTM (Long Short Term Memory)

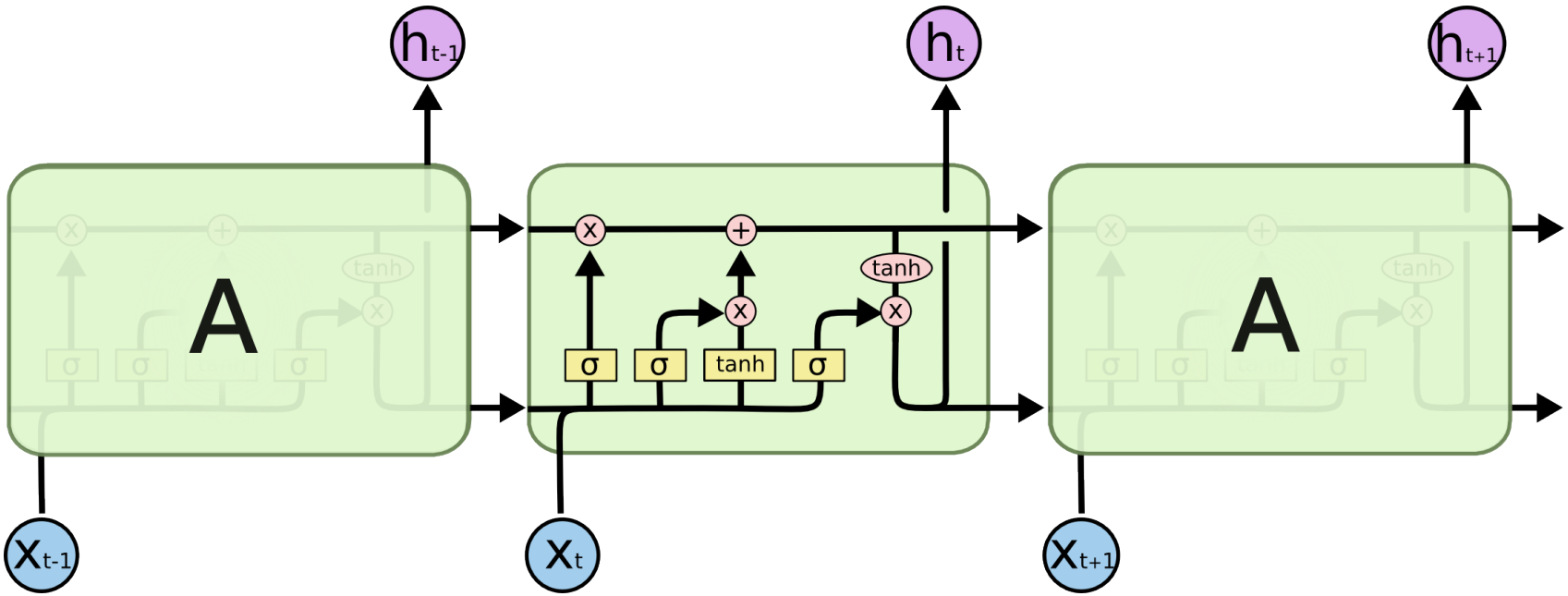


Figura: Understanding LSTM Networks, Christopher Colah 2015.

Comportamiento de las redes sociales

- Durante el proceso de paz se presenció una alta polarización en las redes sociales.

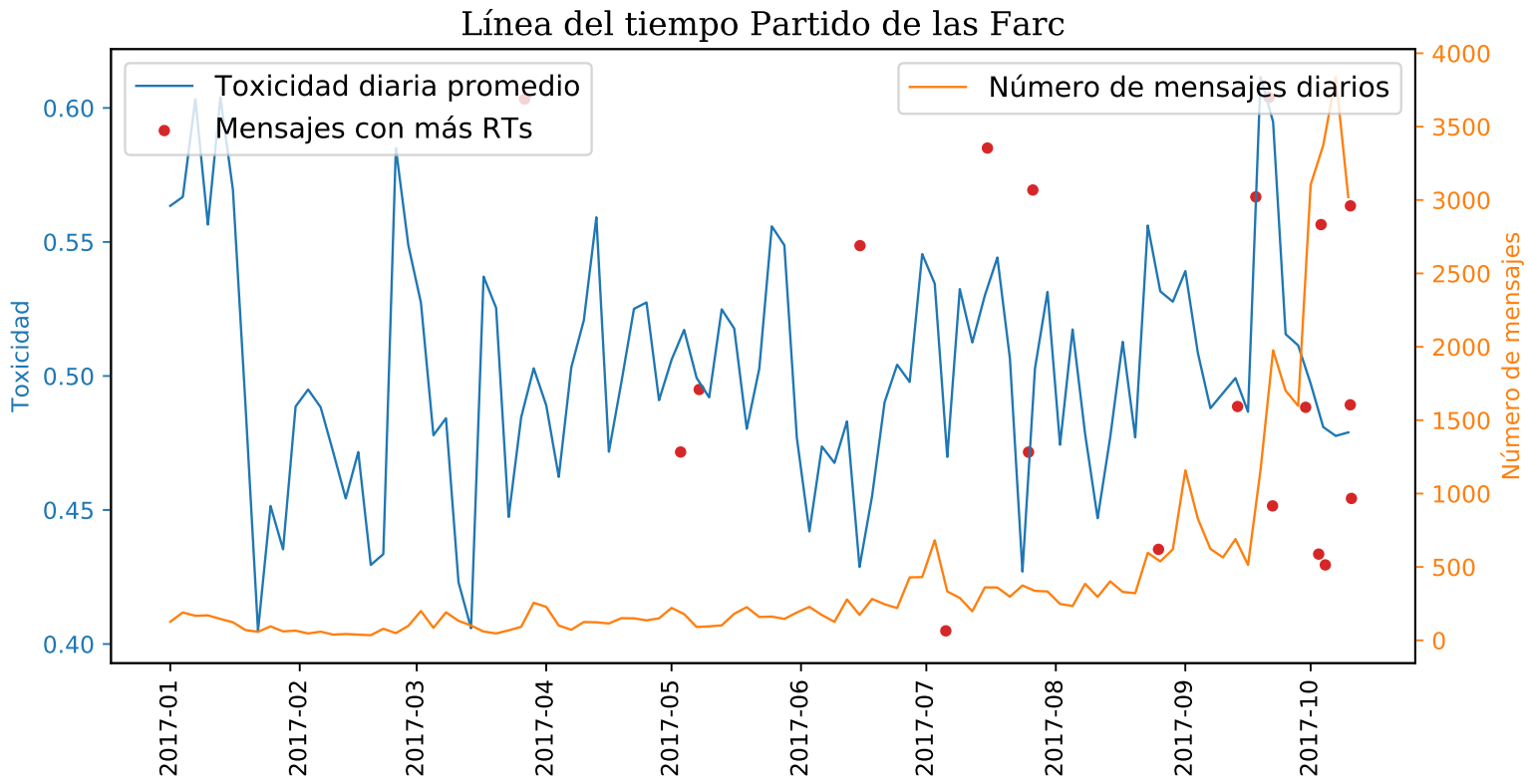
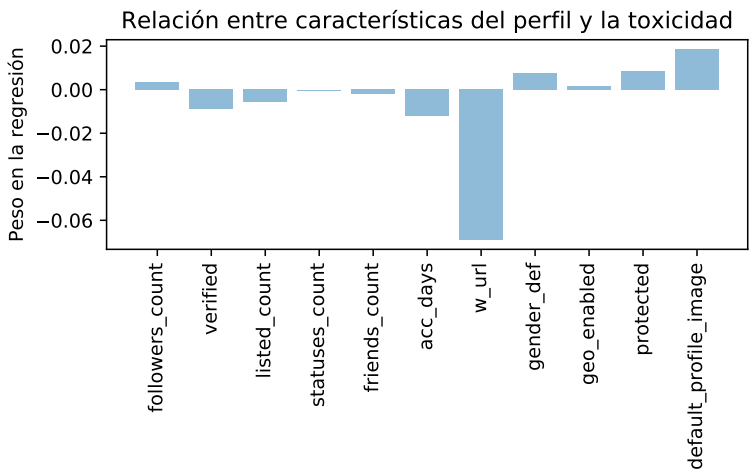
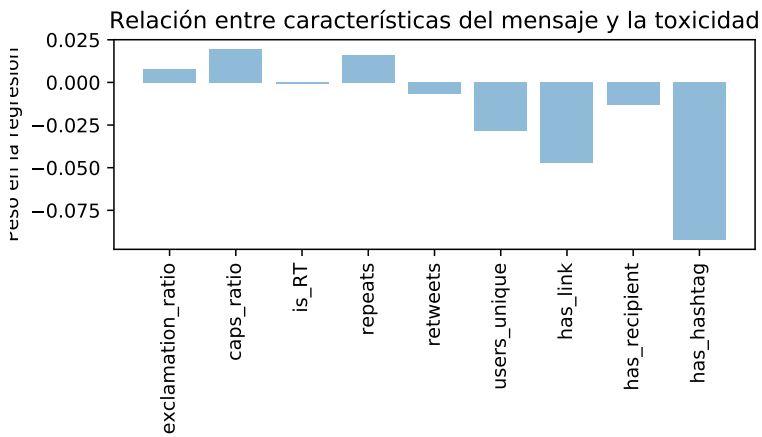


Figura: Linea de tiempo de tweets y su toxicidad.

Análisis de sentimiento

Utilizando Algoritmos de clasificación como regresión logística o árboles aleatorios se identifican los mensajes tóxicos y sus características. Lo mismo se puede hacer para los perfiles de los usuarios.





Identificación de grupos claves

cluster	count	followers_count mean	tweets_day mean	verified mean	msg_count mean	mean_repeats mean	mean_unique mean	retweets mean	score
0	33280	2,398.4	5.6	0.0	1.5	1.1	0.3	1.0	0.0
1	36	9,008.6	25.5	0.0	9.9	134.0	18.5	0.1	0.5
2	657	196,167.1	24.2	1.0	2.2	1.6	0.6	15.0	0.0
3	26	122,647.2	930.2	0.2	4.9	7.2	4.0	1.3	1.0
4	27	248,205.2	13.1	0.6	3.9	3.3	2.0	507.5	0.0
5	17	7,320.0	128.0	0.0	151.9	1.9	0.5	1.2	0.5
6	37	3,917,430.7	160.9	1.0	8.8	4.7	2.6	36.8	0.0
7	433	29,281.2	157.2	0.0	3.5	3.1	1.9	2.0	0.5
8	64	937.6	5.0	0.0	1.2	109.2	103.9	1.2	0.0
9	517	2,392.9	22.3	0.0	23.5	1.7	0.2	2.2	0.0
10	1	16,237,961.0	40.6	1.0	1.0	3.0	3.0	53.0	0.0
11	343	2,552.6	25.6	0.0	1.4	27.5	26.4	0.5	0.5
12	133	87,865.7	16.1	0.3	3.8	1.6	0.8	171.1	0.0

Cuadro: Identificación de bots

cluster	count	followers_count mean	verified mean	friends_count mean	following mean	listed_count mean	retweets mean	score
0	34641	2,421.1	0.0	759.2	0.0	18.8	1.0	0
1	50	1,638,087.4	0.9	5,776.8	0.0	5,367.6	25.2	1
2	82	276,594.6	0.3	2,439.2	1.0	603.2	29.7	1
3	608	124,496.3	1.0	2,523.2	0.0	515.9	13.9	1
4	19	4,846,204.2	1.0	19,734.2	0.4	14,794.1	45.9	1
5	3	1,696,799.0	0.7	673,827.7	0.0	4,005.7	3.6	0
6	116	85,662.6	0.3	3,333.3	0.0	305.6	191.6	1
7	2	11,346,163.0	1.0	798.0	0.0	52,453.5	36.9	1
8	20	300,493.6	0.6	2,144.7	0.1	890.5	560.3	1
9	30	226,817.3	0.1	109,075.9	0.0	1,243.8	24.0	0

Cuadro: Identificación de famosos

Análisis de redes

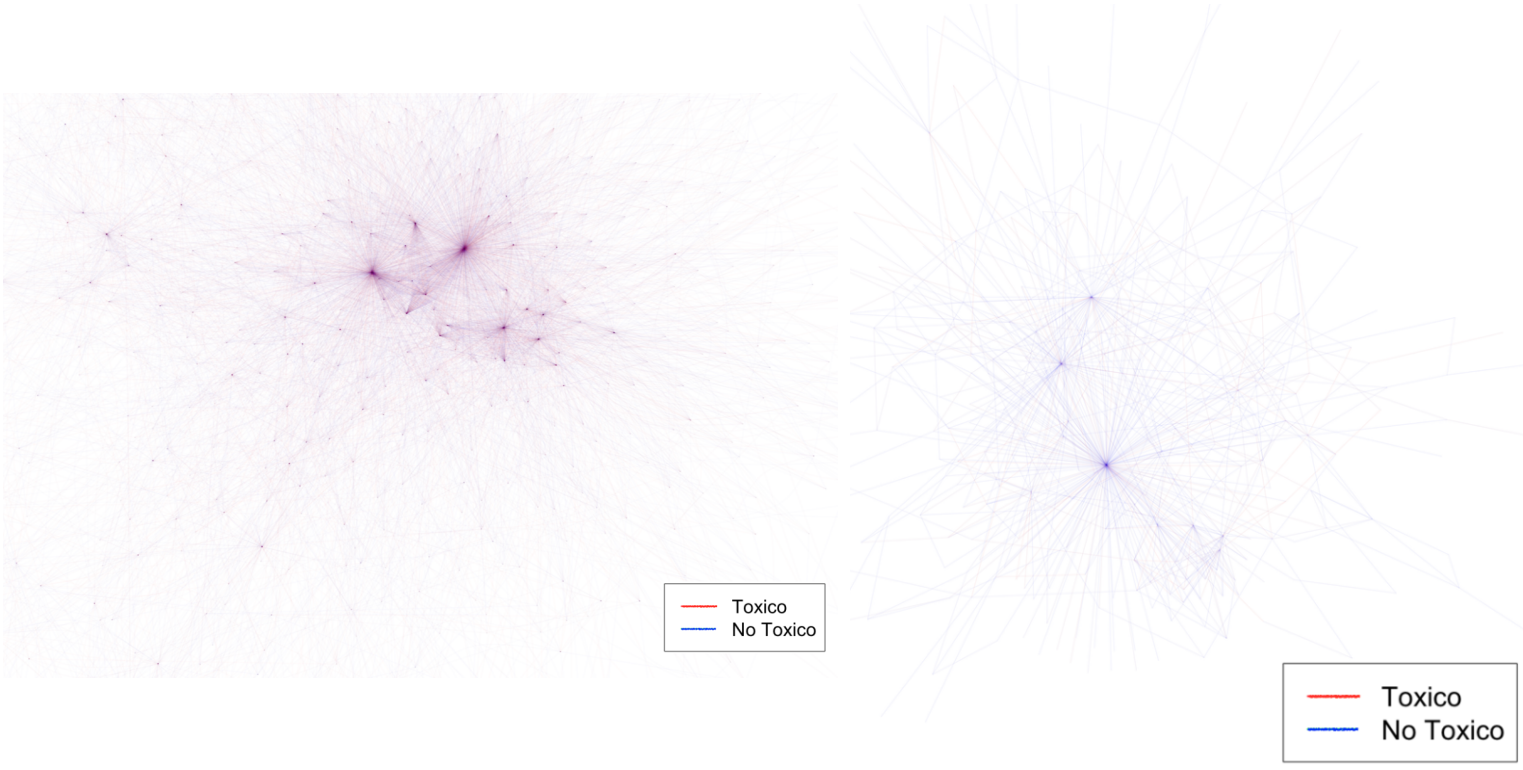


Figura: Redes de conversaciones en Twitter para dos casos. Izquierda: Partido de las Farc. Derecha: Genero (caso Andrea Guerrero.)

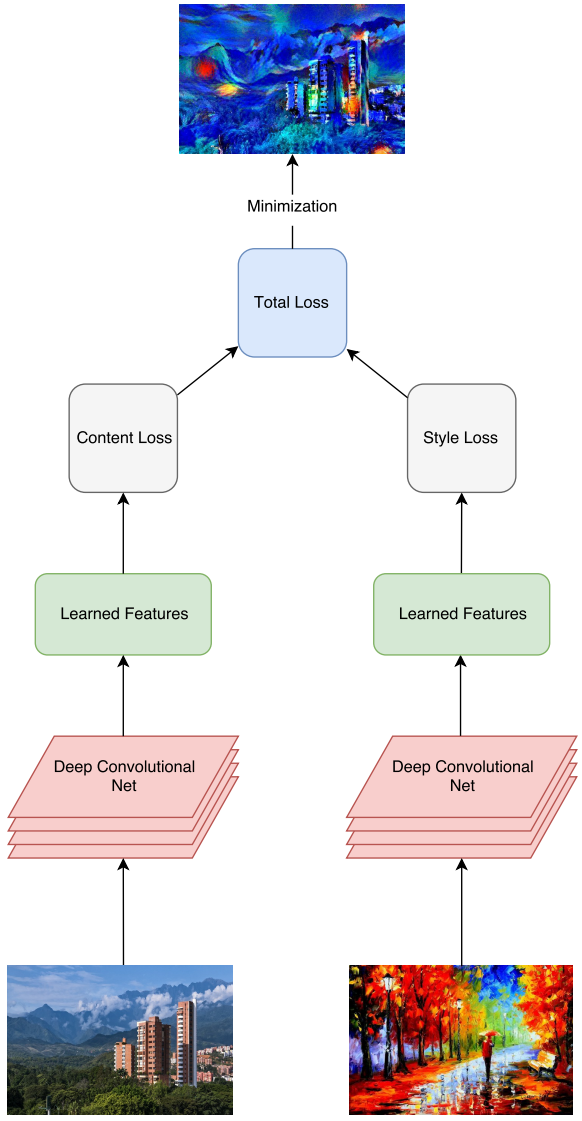
Arte con redes neuronales profundas

Figura: Imágen de contenido

Figura: Imágen de estilo

Resultado

Arte con redes neuronales profundas



GRACIAS