SciPy LATAM·2019 7ª Conferencia Latinoamericana de Python Científico

Extracción y análisis de información de accidentes de tránsito desde redes sociales

Néstor Suat-Rojas
nestor.suat@aunarvillavicencio.edu.co
Profesor Universitario
AUNAR Villavicencio

Octubre 8 al 10 de 2019
Bogotá, Colombia. Universidad de Los Andes.

Corporación Universitaria Autónoma de Nariño
Código SNIES 3817. Resolución 6344 del 17 de Octubre de 2006 MEN.
Estudia por Pasión!

ISO 9001: 2015
BUREAU VERITAS Certification
No. CO18.01029

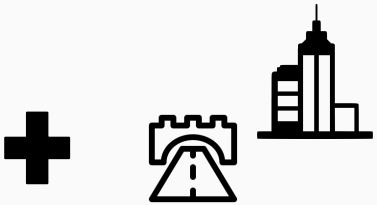
CONTENIDO

- Motivación
- Extracción de accidentes de tránsito en redes sociales
- Resultados

MOTIVACIÓN

Crecimiento poblacional

Crecimiento Económico



Misma Infraestructura

Congestión Vial

Ciudades más congestionadas

(1) Moscú

(2) Estambul

(3) Bogotá

(4) Ciudad de México

(5) São Paulo

* INRIX 2018 Global Traffic

* Cookson y Pishue, 2017 INRIX Global Traffic Scorecard.

MOTIVACIÓN

ACCIDENTALIDAD DE TRÁNSITO

Personas fallecidas **7.000**

40.000 Personas lesionadas

Miles de personas mueren o resultan heridas en accidentes de tránsito cada año.

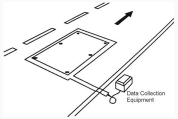
(Observatorio Nacional de seguridad vial, 2017)

VISION ZERO

NO MORE TRAFFIC DEATHS

MOTIVACIÓN

MONITOREO DE TRÁFICO

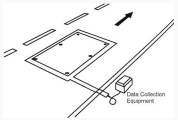


Desafíos encontrados

* Costos y mantenimiento
* Ubicación fija para la recolección
* Cobertura a calles principales
* Pierde exactitud con climas adversos
* Errores de precisión

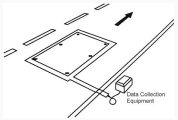
MOTIVACIÓN

MONITOREO DE TRÁFICO



MOTIVACIÓN

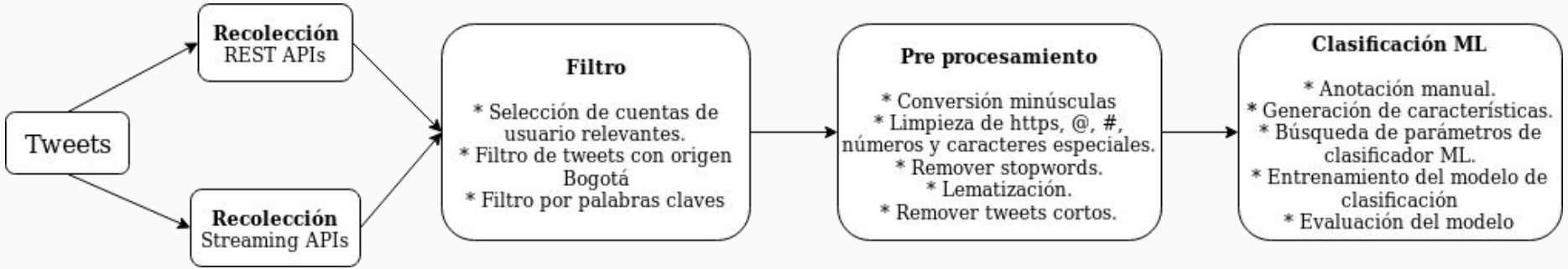
MONITOREO DE TRÁFICO



¿Cómo extraer información de las redes sociales relacionado con accidentes de tránsito en Bogotá?

Método de detección de accidentes de tránsito

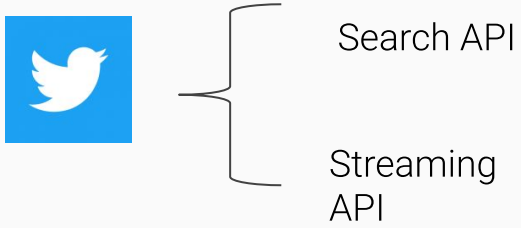
Clasificación automática de accidentes de tránsito en Twitter



* (Schulz et al. 2013, Wang et al. 2015, Gu et al. 2016, Nguyen et al. 2016, Gal-Tzur et al. 2015 & 2017, Salas et al. 2017 & 2018, Zhang et al. 2018)

Recolección

OCT a DIC 2018



Timeline User
@BogotaTransito, @Citytv, @RedapBogota, @WazeTrafficBog, @CIVICOSBOG, @rutassitp, @SectorMovilidad, @UMVbogota, @idubogota, @transmilenio, @IDIGER.

Palabras claves

- ("accidente" OR "choque" OR "incidente vial" OR "incidente" OR "choque entre") -RT -"plan de choque"
- ("atropello" OR "tráfico" OR "trafico" OR "tránsito" OR "transito" OR "#trafico" OR "#traficobogota" OR "sitp" OR "transmilenio") -RT

Geolocalización Bogotá
Coordenada del centro y radio de cobertura.

`Tweepy y PyMongo (MongoDB)`

Método de detección de accidentes de tránsito

Recolección

OCT a DIC 2018

Search API

Streaming API

Bogotá
1’115.378 tweets

Keywords
87.773 tweets

Usuarios y menciones
198.134 tweets

1’401.285

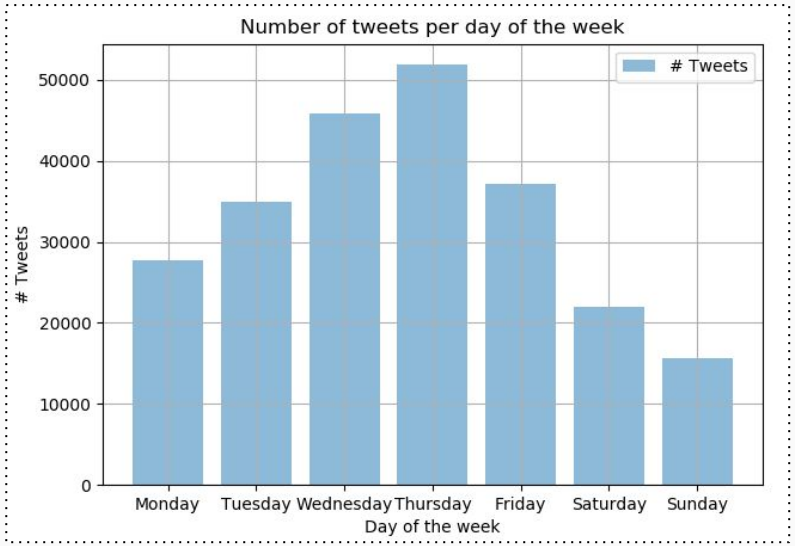
created_at
text
geo
coordinates
place_type
place_name
place_country
place_coordinates
retweet_count

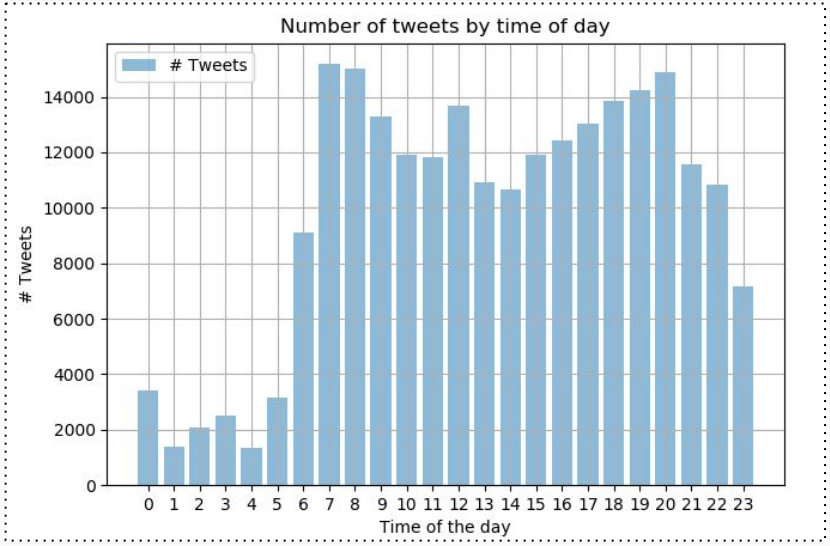
1.72%
tweets
(Poseen georreferencia)

Método de detección de accidentes de tránsito

Recolección

OCT a DIC 2018





Preprocesamiento

"Un motero golpea a una señor con la moto en la carrera 144 con calle 143 en Bogotá @BogotaTransito https://t.co/70sQl2ObH."

spaCy

Tokenización y lowercase

1

un motero golpea a una señor con la moto en la carrera 144 con calle 143 en bogotá @BogotaTransito https://t.co/70sQl2ObH

Eliminar http y @

2

un motero golpea a una señor con la moto en la carrera con calle en bogotá

Eliminar stopwords y lematización

3

output:

```
motero golpear a señor moto
carrera   calle     bogotá
```

Generación de características

TF-IDF

motero golpear a señor
accidente entre motero
atropella a señor

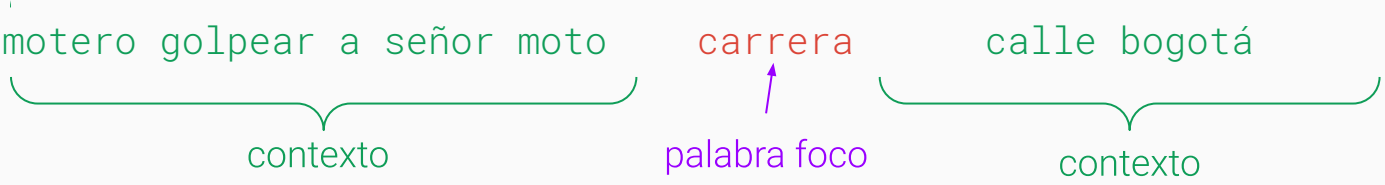
~1000 columnas

accidente	accidente entre	atropella	atropella señor	golpear	...
0	0	0	0	0.577	...
0.5	0.5	0	0	0	...
0	0	0.707	0.707	0	...

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(ngram_range=(1,2), max_df=0.45, min_df=0.001, max_features=None)
features = tfidf.fit_transform(texts)
```

Generación de características

DOC2VEC



DBOW (200 columnas)
Distributed Bag of Words

✚

DM (200 columnas)
Distributed Memory

~400 columnas

	X1	X2	X3
motero golpear a señor	0.6	0.2	0.2
accidente entre motero	0.75	0.24	0.1
atropella a señor	0.1	0.142	0.7

Generación de características

TF-IDF

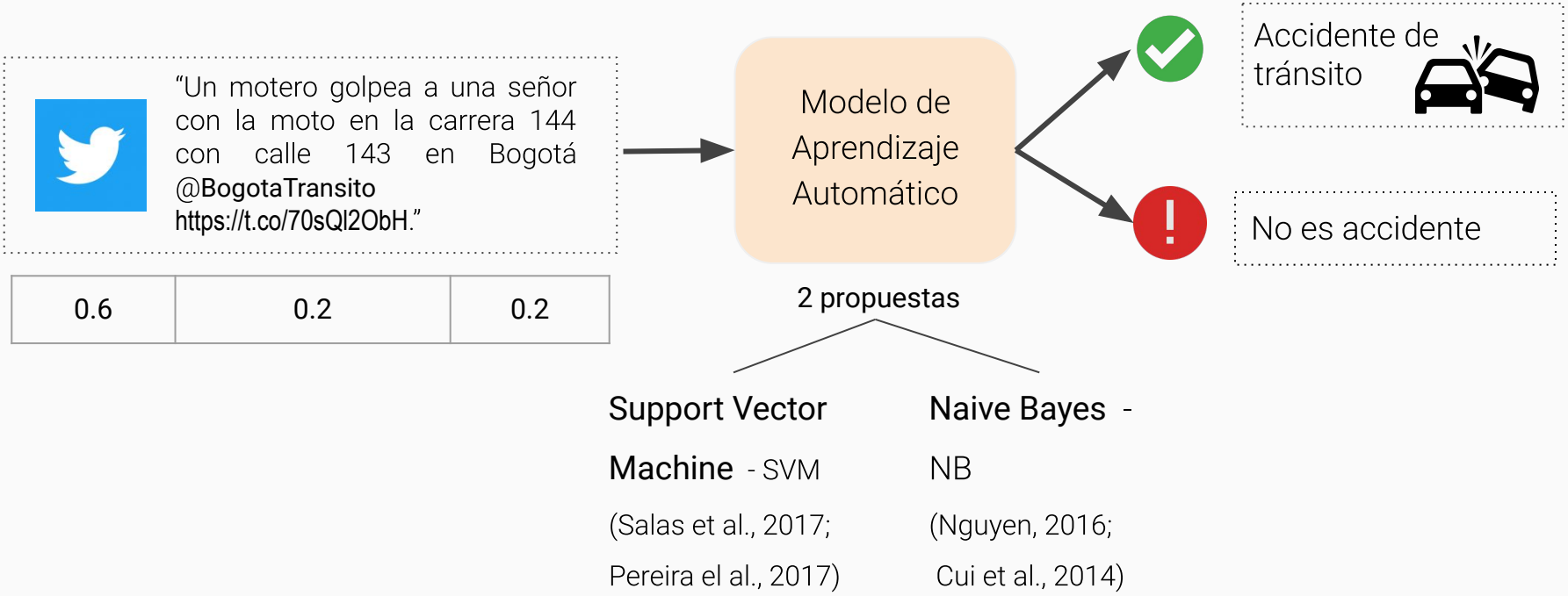
```
TfidfVectorizer(
ngram_range=(1,1), max_df=0.3,
min_df=0.001, max_features=1000
)
```

DOC2VEC

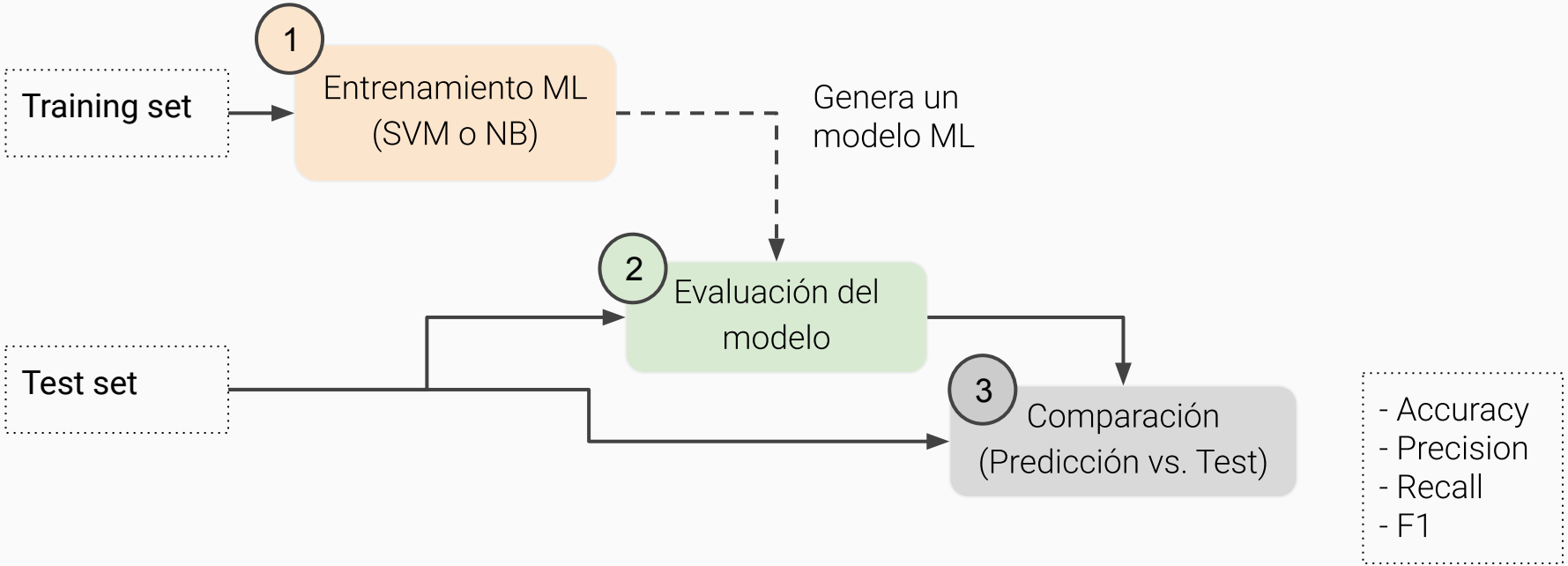
```
from gensim.models.doc2vec import Doc2Vec

ddm = Doc2Vec(vector_size=200,
window=5, alpha=0.025,
min_alpha=0.0001, min_count=5,
dm=1, dm_mean=1, epochs=40
)

dbow = Doc2Vec(vector_size=200,
window=5, alpha=0.025, min_alpha=0.0001,
min_count=5, dm=0, epochs=40
)
```

(Pereira et al., 2017; Le & Mikolov, 2014)

Clasificación automática con ML



Entrenamiento del modelo ML Supervisado



Etiquetado

Fase 1.
15.000 tweets, de los cuales 1941 positivos y 10944 negativos.
13 personas participaron, tomó una semana

Fase 2.
7582 tweets, de los cuales 723 positivos y 6494 negativos.
20 personas participaron y tomó 3 semanas

# Tweets positivos	# Tweets negativos	Total
2664	2664	5328

Muestras

Relacionado a accidente	A los que van o piensan ir por la Carrera 30: Se reporta un accidente en el sentido sur-norte, al parecer, con un fallecido. El tráfico es porque tienen acordonada la zona para realizar levantamiento. Tomen vías alternas. #Carrera30 #Accidente
Relacionado a accidente	Incidente vial entre bus 🚍 y un motociclista 🏍️ en la calle 86a con carrera 111a. Unidad de 👮@TransitoBta y 🚑 asignadas.
Errores de ortografía	@TransMilenio por favor enviar bises la calle 80 esta colapsada por el accidente de la.avda cali.... estamos desde las 7 y no se puede ingresar por la gran cantidad de gente
Dos reportes en el mismo tweet.	en la avenida Primero de Mayo con carrera 69 en sentido occidente - oriente chocan un taxi y una motocicletaen la avenida de La Esperanza con carrera 68 A en sentido occidente - oriente chocan un vehículo particular y una camioneta

Búsqueda de parámetros de SVM y TFIDF

```
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report
```

DATASET

TRAIN / VALIDATION	TEST

- GridSearchCV (cv=5)	- Scores
	  - Accuracy
	  - F1

TFIDF

Best score: 0.956

max_df: 0.3 max_features: 1000 min_df: 0.001

ngram_range: (1, 1)

SVM

Best score: 0.945

C: 4 gamma: 0.7 kernel: 'rbf'

Resultados

Train 3729 **Test** 1599 **Total** 5328

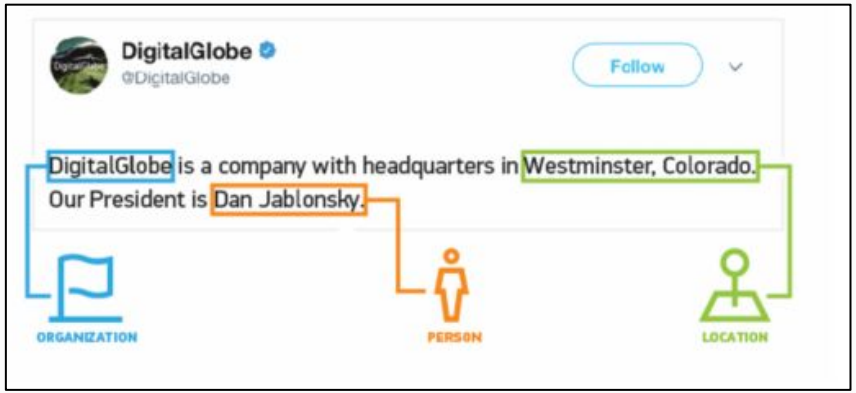
Embedding	Clasificador	Accuracy	F1	Precision	Recall
TFIDF	SVM C: 4 gamma: 0.7 kernel: 'rbf'	**0.969356**	**0.968968**	**0.970812**	**0.96713**
	NB	0.854909	0.866667	0.794521	0.953224
Docv2vec	SVM C: 0.1 gamma: 0 kernel: 'linear'	0.912445	0.90991	0.926606	0.893805
	NB	0.819262	0.793719	0.911475	0.702908

Conclusiones

- TF IDF mejor desempeño como línea de base (**96%** de exactitud). Al igual que requiere menor tiempo de entrenamiento.
- Doc2vec requiere de un corpus grande para su entrenamiento.
- Support Vector Machine es un modelo rápido de entrenar con resultados similares a otros modelos.
- Para la ciudad de Bogotá se ha diseñado un modelo que sirve de línea base, demostrando información disponible sobre accidentes de tránsito y un mecanismo viable para su extracción.

Trabajos Futuros

- Detección de entidades nombradas y Geoparsing.



- Fusionar los datos recolectados en redes sociales con otras técnicas de monitoreo.

Bibliografía

1. Nguyen, H., Liu, W., Rivera, P., & Chen, F. (2016). TrafficWatch: Real-Time Traffic Incident Detection and Monitoring Using Social Media.
2. Schulz, A., Ristoski, P., & Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs.
3. Gutiérrez, C., Figueiras, P., Oliveira, P., Costa, R., & Jardim-goncalves, R. (2016). An Approach for Detecting Traffic Events Using Social Media.
4. Gu, Y., Qian, Z. (Sean), & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data.
5. Caimmi, B., & Vallejos, S. (2016). Geolocalización de incidentes de tránsito a partir del análisis de sentencias extraídas de redes sociales. Universidad Nacional Del Centro de La Provincia de Buenos Aires.
6. Salas, A., Panagiotis Georgakis, Y. P. (2018). Incident Detection Using Data from Social Media.
7. Zhang, Z., He, Q., Gao, J., & Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data.
8. Pereira, J., Pasquali, A., Saleiro, P., & Rossetti, R. (2013). Transportation in Social Media: An Automatic Classifier for Travel-Related Tweets.

Néstor Suat-Rojas

nestor.suat@aunarvillavicencio.edu.co

Ingeniero de Sistemas
Profesor Universitario
AUNAR Villavicencio