# Identificación de la Oferta y Demanda de Medicinas en Twitter

L. MARTÍNEZ, Universidad Simón Bolívar.
N. MAÑAN, Universidad Simón Bolívar.
J. RIVAS, Universidad Simón Bolívar.

---

## RESUMEN

La crisis de medicinas en Venezuela ha llevado a la población a usar *Twitter* como herramienta para buscar y ofrecer medicamentos. Este estudio busca facilitar la búsqueda de medicamentos en *Twitter* usando un algoritmo de inteligencia artificial bajo aprendizaje supervisado. Se entrenó una red neuronal con un set de entrenamiento y pruebas de 6000 tuits clasificados manualmente y pre-procesados según características que experimentalmente dieron mejor resultado en la clasificación. Se obtuvo como resultado una red neuronal capaz de identificar con una precisión del 84% y una exhaustividad de 92% para tuits de oferta, 87% para demanda y 74% para otros tuits. La implementación y entrenamiento de una red neuronal para la identificación automática de tuits puede ser de ayuda importante para la localización de medicamentos.

---

## AGRADECIMIENTOS

A la Profa. Carolina Martínez por la concepción de la idea de este estudio así como su apoyo durante el desarrollo del mismo.

---

## 1 INTRODUCCIÓN

Un hecho actual al que muchas personas se enfrenta es a la escasez de medicamentos y la dificultad de conseguir los mismos en farmacias, clínicas y otros entes de salud pública, debido a la actual demanda surge la necesidad de utilizar las redes sociales como vía para comunicar y hacer saber de la existencia de un medicamento o para solicitar el mismo. *Twitter* es una de las redes donde inclusive existen cuentas dedicadas a difundir las distintas solicitudes, sin embargo, persiste la dificultad de encontrar personas que estén ofertando cierto medicamento solicitado o cuya demanda pase inadvertida ante el gran flujo de tuits de usuarios con distintas urgencias.

El siguiente estudio plantea como solución la implementación y entrenamiento de un algoritmo de aprendizaje supervisado, en este caso una red neuronal, que discierne si un tuit es de oferta o de demanda de medicamentos e identifica el medicamento en cuestión, con el objetivo de facilitar la localización de medicinas para los usuarios que utilizan *Twitter* para estos fines.

---

## 2 MARCO TEÓRICO

### 2.1 Red Neuronal

Modelo computacional basado en un conjunto de neuronas artificiales que simulan el comportamiento de las de las neuronas en los cerebros biológicos.

## 2.2 Validación Cruzada (*Crossvalidation*)

Técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

## 2.3 API (*Application Programming Interface*)

Conjunto de subrutinas, funciones y procedimientos que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción [2].

## 2.4 Marco de Trabajo (*Framework*)

Estructura conceptual y tecnológica de soporte definido que puede servir de base para la organización y desarrollo de software [3].

---

# 3 DISEÑO DE LA SOLUCIÓN

## 3.1 Recolección De Datos

Para los algoritmos de aprendizaje se necesitan un conjunto de datos iniciales para su entrenamiento, para este caso no se disponía de datos ya procesados por lo que vio la obligación de recolectar y filtrar tuits para entrenar el algoritmo.

Para la recolección se disponía de la herramienta *Birdwatcher* [4], un marco de trabajo de OSINT, fuente de inteligencia libre (*Open-Source Intelligent*), para *Twitter,* que facilita la recolección de tuits de varias cuentas de usuarios. Para esto se necesitaba crear una API en *Twitter,* para poder acceder y leer a los datos de la red social.

Con la ayuda de esta herramienta se recolectaron tuits de 33 cuentas relacionadas con la solicitud de de medicamentos y afines para obtener un total de 24.745 tuits

## 3.2 Preprocesamiento De Datos

Una vez recolectado los tuits de las distintas cuentas, se realizó un primer filtro a través de un *script* que buscara la raíz de ciertas palabras claves y otras palabras específicas para obtener tuits referentes a la oferta y demanda de medicamentos. La lista de palabras puede verse en el Anexo 1, de este primer filtro se redujo la cantidad de tuits a 16.517.

Para obtener una muestra útil para el entrenamiento del algoritmo se dividieron los datos en ejemplos positivos y negativos para de esta manera lograr discernir los resultados que se quieren obtener, para este estudio era necesario clasificar la muestra total en tuits de demanda, oferta y casos nulos, donde el tuit no hacía referencia a ninguno. Además, se toma en cuenta solo casos donde nombraban expresamente el nombre del medicamento, como muchos de los tuits llegaban a ser subjetivos para clasificar en casos de oferta o demandas y para evitar falsos negativos o positivos al filtrar a través de un *script,* se etiquetaron manualmente los datos en los tres casos mencionados anteriormente quedando 2.346 tuits de oferta, 4.136 de demanda y 10.035 que no correspondían a las otras categorías que se denominará como “otros”.

Se usaron ciertos criterios para el filtro manual de los datos, de esta manera tuits para las distintas categorías tendrían características similares que permitirán el buen entrenamiento del algoritmo, entre estos se encuentran:

- Tuits que solo mencionan el nombre del medicamento son calificados como oferta del mismo.

- Si se expresa la idea de un trueque o cambio de medicinas se considera como un tuit de oferta.
- Los suplementos son considerados como medicamentos para tratar enfermedades.
- La solicitud y disponibilidad de instrumentos quirúrgicos y otros materiales necesarios para tratamientos (muletas, sillas de rueda, jeringas, catéter, entre otros) no son tomados en cuenta para el conjunto de datos.
- Tuits con información de la disponibilidad del medicamento en alguna farmacia o establecimiento parecido son considerados como oferta del mismo.

Finalmente se ajustó la muestra obtenida para obtener mejores resultados al entrenar el algoritmo y tener un conjunto homogéneo de tuits, por lo que se redujo la cantidad a 2000 tuits de cada categoría.

Además, de los datos finales elegidos se tomaron las primeras 90 medicinas que aparecían más frecuentemente para su identificación entre los tuits, en la siguiente tabla se muestran las medicinas escogidas.

**Tabla 1. Lista de medicinas conocidas según la frecuencia de repeticiones.**

| | | | | | |
|---|---|---|---|---|---|
| Clexane | Sertralina | Ampicilina | Xarelto | Atorvastatina | Trozolet |
| Fenobarbital | Stalevo | Budecort | Avastin | Diclofenac | Bacipro |
| Valpron | Moderan | Ceftriaxona | Cisplatino | Furosemida | Carvedilol |
| Carboplatino | Granocyte | Pegyt | Valsartan | Metfor | Fulgram |
| Epamin | Euthyrox | Ridal | Leucovorina | Metformina | Erbitux |
| Losartan | Plavix | Atenolol | Clindamicina | Meropenem | Tegretol |
| Doxorrubicina | Aciclovir | Neurixa | Keppra | Glucofage | Somazina |
| Concor | Glaucotensil | Ulcon | Plaquinol | Captopril | Trileptal |
| Albumina Humana | Pradaxa | Aldactazida | Unasyn | Dexametasona | Carbamazepina |
| Valcote | Oxicodal | Clonazepam | Madopar | Azitromicina | Enalapril |
| Benicar | Femara | Bicalutamida | Amlodipina | Amikacina | Levofloxacina |
| Pregabalina | Dostinex | Pentoxifilina | Tramal | Inmunoglobulina | Alprazolam |
| Prednisona | Insulina | Humalox | Neupogen | Aluron | Herceptin |
| Sinemet | Badan | Novolin | Hidroclorotiazida | Methotrexato | Digoxina |
| Metronidazol | Omeprazol | Solumedrol | Aprovel | Ciprofloxacina | Acido Valproico |

## 3.3 Selección de Atributos

Los tuits recolectados por *Birdwatcher* tienen un formato prestablecido, llevan atributos asociados por defecto, para el fin de este proyecto y el entrenamiento del algoritmo se seleccionaron ciertos atributos del formato establecido y se agregaron otros que se consideraron necesarios como el TF-IDF [5] frecuencia de término-frecuencia inversa de documento (*Term Frequency – Inverse Document Frequency*):

• **Length:** largo, en número de caracteres, del tuit.
• **Retweet_count:** Número de veces que se reenvía un tuit.
• **Hashtags**: Número de las etiquetas para la identificación de ciertas palabras elegidas por el usuario.
• **Contains_url:** Número que indica si el tuit posee enlaces a otras páginas (1) o no (0).
• **Demand_words:** Número de palabras encontradas en el tuit clasificadas como palabras de demanda.
• **Offer_words:** Número de encontradas en el tuit clasificadas como palabras de oferta.
• **Tf-idf_demand:** La suma de tf-idf para cada palabra de demanda del tuit.
• **Tf-idf_offer:** La suma de tf-idf para cada palabra de oferta del tuit.
• **Neg_words:** Número de palabras encontradas en el tuit que pertenecen a la lista de palabras negativas.
• **Pos_words:** Número de palabras encontradas en el tuit que pertenecen a la lista de palabras positivas.
• **N_known_medicines:** Número de medicinas conocidas, almacenadas previamente según su frecuencia en el conjunto de datos de entrenamiento.
• **Cluster:** Número, valor a predecir: oferta (1), demanda (2) u otros (0).

Las tablas 2 y 3 muestran las palabras relacionadas con oferta y demanda con el cálculo de su TF-IDF respectivo.

**Tabla 2. Palabras relacionadas con oferta, junto con su TF-IDF de acuerdo al set de entrenamiento.**

| Palabra | TF | IDF | TF*IDF |
|---|---|---|---|
| **confirmar** | 0.0034785559750562084 | 24.390243902439025 | 0.08484282865990753 |
| **cambiar** | 0.001442328087218428 | 58.8235294117647 | 0.08484282865990751 |

| | | | |
|---|---|---|---|
| **farmacia** | 0.0055996266915538965 | 15.873015873015873 | 0.08888296335799836 |
| **vender** | 0.0005514783862893989 | 166.66666666666666 | 0.09191306438156648 |
| **tener** | 0.010011453781869087 | 8.583690987124463 | 0.0859352255954428 |
| **necesitar** | 0.0039451915326857 | 21.50537634408602 | 0.08484282865990753 |
| **ofrecer** | 0.000890849700929029 | 95.23809523809524 | 0.08484282865990753 |
| **disponibilidad** | 0.052348025283162944 | 1.6339869281045751 | 0.08553598902477605 |
| **donar** | 0.04242141432995376 | 2.100840336134454 | 0.089120618340239 |
| **conseguir** | 0.0002121070716497688 | 400.0 | 0.08484282865990753 |

**Tabla 3. Palabras relacionadas con demanda, junto con su TF-IDF de acuerdo al set de entrenamiento.**

| Palabra | TF | IDF | TF*IDF |
|---|---|---|---|
| **informar** | 0.003854239663629993, | 20.408163265306122 | 0.07865795231897946 |
| **solicitar** | 0.013781826676010278 | 5.681818181818182 | 0.07830583338642202 |
| **buscar** | 0.020711671727789458 | 4.842615012106537 | 0.10029865243481578 |
| **agradecer** | 0.0038931713774040333 | 20.0 | 0.07786342754808066 |
| **comprar** | 0.0006229074203846453 | 125.0 | 0.07786342754808066 |
| **tratamiento** | 0.002491629681538581 | 31.25 | 0.07786342754808066 |
| **adquirir** | 7.786342754808066e-05 | 1000.0 | 0.07786342754808066 |
| **urgente** | 0.0444989488437281 | 1.8083182640144666 | 0.08046826192355895 |
| **favor** | 0.01164058241843806 | 6.8728522336769755 | 0.08000400287586296 |
| **necesitar** | 0.029938487892237017 | 2.638522427440633 | 0.07899337174732722 |
| **donde** | 0.00825352332009655 | 9.47867298578199 | 0.07823244853172086 |
| **gracias** | 0.00817565989254847 | 9.523809523809524 | 0.07786342754808066 |
| **ayudar** | 0.004593942225336759 | 17.094017094017094 | 0.07852892692883348 |
| **tener** | 0.0021801759713462585 | 35.714285714285715 | 0.07786342754808066 |
| **emergencia** | 0.00019465856887020166 | 400.0 | 0.07786342754808066 |
| **requerir** | 0.00817565989254847 | 9.569377990430622 | 0.07823597983299971 |
| **conseguir** | 0.004165693373822316 | 18.69158878504673 | 0.07786342754808068 |

## 3.4 Entrenamiento del Algoritmo

Se decidió implementar una red neuronal para el procesamiento de los datos, la red neuronal está diseñada con una sola capa oculta y tres neuronas en la capa de salida, se probaron diferentes números de neuronas en la capa oculta para entrenar la red, se obtuvo mejores resultados con 10, 15 y 20 neuronas y 150 épocas (iteraciones), para el entrenamiento. Se dividieron los datos homogéneamente en porcentajes 90-10 usando cross-validation para seleccionar el mejor conjunto de datos.

---

# 4 RESULTADOS OBTENIDOS

De la partición realizada para el *cross-validation* se obtuvo que el mejor conjunto de datos fue el conjunto cero (0), en la siguiente tabla se pueden ver los valores obtenidos para 10, 15 y 20 neuronas y 150 épocas aplicadas a este conjunto. Además se incluye dentro de los datos recolectados el resto de conjuntos de entrenamiento**.**

**Tabla 4. Comparación de neuronas con el conjunto cero.**

| **Neuronas** | 10 | 15 | 20 |
|---|---|---|---|
| **MSE (Mínimo Error Cuadrático)** | 0.343333 | 0.36 | 0.368333 |
| **Precisión** | 0.846667 | 0.84 | 0.841667 |
| **Índice de Verdaderos Positivos (Oferta)** | 0.925 | 0.905 | 0.905 |
| **Índice de Verdaderos Positivos (Demanda)** | 0.87 | 0.84 | 0.87 |

| | | | |
|---|---|---|---|
| **Índice de Verdaderos Positivos (Otros)** | 0.745 | 0.775 | 0.75 |
| **Porcentaje de valores positivos predichos (Oferta)** | 0.825893 | 0.830275 | 0.841860 |
| **Porcentaje de valores positivos predichos (Demanda)** | 0.84878 | 0.848485 | 0.824645 |
| **Porcentaje de valores positivos predichos (Otros)** | 0.871345 | 0.842391 | 0.862069 |

Se puede notar que, a pesar de pequeñas diferencias, el conjunto con 10 neuronas muestra mayor índice para clasificar los tuits en las categorías mencionadas.

A modo de prueba, se realizó un pequeño script usando la API de *Twitter* para recolectar tuits en tiempo real y procesarlos por la red neural diseñada para automáticamente clasificar los tuits según lo explicado anteriormente y ejercer el propósito del proyecto Medicinas en Twitter.

---

## 5 CONCLUSIONES

Con este estudio se concluye que con una adecuada selección de atributos es posible clasificar con una presicion aceptable tuits de demandas y ofertas de medicamentos en *Twitter*.

Es importante tener en cuenta que los tuits fueron clasificados de manera manual, siendo posible que algunas instancias del conjunto de entrenamiento esten incorrectamente etiquetadas y, por lo tanto, es posible aún mejorar los resultados obtenidos con otros sets de datos con mayor precisión en la clasificación.

Existe también un potencial de uso importante para este tipo de clasificación, en el futuro se podría utilizar este estudio para desarrollar una aplicación que permita el enlace de personas que buscan un medicamento con aquellas que lo ofrezcan.

---

## 6 ANEXOS

**Anexo 1:** Lista de palabras claves y raíces de palabras para el primer filtro de tuits recolectados:

*"solicit", "necesit", "urgen", "medic", "mg", "gr", "dosis", "ampollas", "pastillas", "disponib", "fundaci", "hosp", "dona", "requi", "reque", "paciente", "serviciopublico", "pediat", "tabletas", "comprimidos", "tratamiento".*

---

## REFERENCIAS


[1] Tom M. Mitchell. 1997. Machine Learning.

[2] Es.wikipedia.org. (2017). *Interfaz de programación de aplicaciones*. [online] Disponible en: https://es.wikipedia.org/wiki/Interfaz_de_programaci%C3%B3n_de_aplicaciones [Accedido 30 Mar. 2017].

[3] Es.wikipedia.org. (2017). *Framework*. [online] Disponible en: https://es.wikipedia.org/wiki/Framework [Accedido 30 Mar. 2017].

[4] GitHub. (2017). *michenriksen/birdwatcher*. [online] Disponible en: https://github.com/michenriksen/birdwatcher [Accedido 30 Mar. 2017].

[5] Es.wikipedia.org. (2017). *Tf-idf*. [online] Disponible en: https://es.wikipedia.org/wiki/Tf-idf [Accedido 30 Mar. 2017].