Facultad de ingeniería

Escuela de Ingeniería de Sistemas

SISTEMA DE DETECCIÓN DE LA CARGA OFENSIVA DE COMENTARIOS EN REDES SOCIALES.

Maximiliano Casale

Andrés Medina

Caracas, noviembre de 2018

# Título tentativo:

# Sistema de detección de la carga ofensiva de comentarios en redes sociales.

# Planteamiento del problema

Las redes sociales se han vuelto parte del día a día de la mayoría de las personas, haciendo que los individuos más populares en las mismas generen un impacto importante en la sociedad, generando grandes oleadas de comentarios a favor o en contra de cada publicación. Por tal razón, las empresas de redes sociales han visto la necesidad de generar un sistema que permita monitorear el comportamiento de los usuarios en los comentarios debido a que ha surgido una creciente tendencia de crear polémicas mediante el uso de insultos o material ofensivo, desmejorando la calidad y satisfacción de los clientes que se entretienen haciendo debates en la sección de comentarios de cada publicación.

Aquellos usuarios que tienen una personalidad violenta o tóxica son un problema para las empresas de redes sociales. Según el psicólogo argentino Stamateas (2008) son aquellas personas “que pareciera encuentran placer en hacernos difícil la convivencia o nuestro trabajo.” (p.49). Estos usuarios ven la necesidad de resaltar en la sociedad mediante el uso de comentarios ofensivos. Debido a esto, algunas empresas de software han optado por la creación de sistemas de prevención y detección de comentarios tóxicos para mejorar la experiencia y la calidad de servicio que reciben los usuarios.

Dentro de la industria de desarrollo de software, *Google* está creando una herramienta de inteligencia artificial para tratar de combatir el acoso en línea y poder determinar si una conversación se está tornando agresiva. Sin embargo, está en una fase de desarrollo temprana y únicamente está disponible en el idioma inglés.

En el caso particular de *Instagram*, no existe un control en cuanto a la toxicidad de los comentarios escritos en español, por lo cual una publicación se puede llegar a minar de comentarios tóxicos, evitando que se pueda disfrutar de un servicio de calidad, disminuyendo los usuarios activos y por ende las ganancias por publicidad y venta de datos en la aplicación.

Deseando mejorar la calidad de servicio de las redes sociales y considerando lo anteriormente expuesto, en el presente trabajo de investigación se propone la creación de una herramienta que permita detectar y clasificar los comentarios ofensivos en español, con el fin de identificar de forma automatizada aquellos mensajes agresivos y brindar a las empresas de redes sociales una alternativa para lidiar con los usuarios de personalidad violenta de habla hispana.

# Delimitación y alcance del tema.

En las empresas que manejan grandes volúmenes de datos se ha hecho necesario el uso de herramientas que automaticen la detección de patrones entre los registros. Para este requerimiento en específico se han desarrollado diferentes técnicas de *Machine Learning* cuya eficacia depende del escenario en que se pone a prueba y la arquitectura que se utilice. Regularmente, para casos de detección o clasificación de patrones se utilizan las denominadas redes neuronales cuya intención es análoga a simular el funcionamiento de las neuronas del cerebro humano, lo que computacionalmente corresponde a “una red interconectada paralelamente con elementos simples que siguen una interacción jerárquica” (Kohonen, 1989, p. 251).

Dentro del ámbito de redes neuronales existen diferentes modelos de arquitecturas de red, los cuales se destacan en diferentes categorías. Por ejemplo, para el análisis de cadenas de caracteres, el modelo *Long Short-Term Memory* (en adelante por sus siglas en inglés: LSTM) es una red neuronal recurrente con una estructura de celdas con compuertas de entrada y salida que permiten recordar y olvidar datos de contexto en series (Gers, Schmidhuber, Cummins, 1999) y es conocida por su capacidad de clasificación y compresión de texto en lenguaje natural y lenguaje sin contexto (Gers, Schmidhuber, s.f.).

Las redes neuronales recurrentes pueden contener ciclos entre los nodos, formando una especie de “memoria” que permite almacenar datos que tengan influencia en la respuesta de la red (Graves, s.f, p.18-19). El modelo LSTM ha demostrado exitosamente que puede resolver problemas de la vida real (como detectar patrones en texto sin contexto) que siguen siendo imposibles para cualquier otro modelo de red neuronal (Graves, s.f., p.32), motivo por el cual se utilizará para la investigación.

Para la introducción de la data en la red, investigaciones anteriores en el área han demostrado con éxito que el uso de N-gramas ha funcionado mucho mejor que especificaciones de la sintaxis del lenguaje o palabras y oraciones embebidas “debido a la mayor robustez que exhiben los N-gramas por las variaciones de ortografía que son muy comunes en discusiones en línea” (Dixon, Thain, Wulczyn, 2017). Por esta razón, para este trabajo de investigación, la red neuronal se entrenará usando registros de texto representados por secuencias de palabras de longitud N, llamadas N-gramas (Bruguier, Damavandi, Kumar, Shazeer, 2016), con la finalidad de detectar intenciones ofensivas en los mensajes o comentarios en español recuperados de la red social *Instagram*, mediante el modelo de red neuronal LSTM mencionado anteriormente.

Para el desarrollo de software en *Machine Learning*, una de las mejores alternativas es *Python* (Rebollo, M. 2018). Este lenguaje es conocido principalmente por sus librerías de acceso público y su fácil y rápida implementación (Python,2018). El beneficio más grande considerado es la velocidad de desarrollo gracias a las librerías que ya han sido desarrolladas e incluyen la mayoría de las funciones necesarias para llevar a cabo un software de inteligencia artificial.

Entre las librerías más populares están: *TensorFlow* (Abadi et al, 2015) (desarrollada por *Google* como librería de *Machine Learning*) y *Keras* (Keras, 2015) (desarrollada por François Challet como *frontend* para *deep learning*). Ambas librerías tienen las herramientas para realizar el modelo de arquitectura planteado y se estarán usando en conjunto durante la investigación.

Para la clasificación de la carga ofensiva de los comentarios, se tomará en consideración el lenguaje de los mismos. Definiendo como lenguaje "la manera de expresarse" (Real Academia Española, 2001). De la misma manera, se tomará una muestra de usuarios que calificarán los comentarios que consideren ofensivos, es decir, si los usuarios que califican detectan alguna de las categorías del lenguaje soez (Blasfemia, Amenazas, Insultos, Dobles sentidos, interjección, Disfemismos, Ironía, sarcasmo, tabúes, entre otros) en los mensajes que van a ser calificados, entonces los marcarán como ofensivo. Dixon, Thain y Wulczyn (2017) recomiendan que la muestra de usuarios sea de alrededor de 40 personas, de forma que se puede tener una base sólida de calificaciones por cada comentario. Estos comentarios calificados se utilizarán para entrenar la inteligencia artificial que determinará la carga ofensiva de nuevos comentarios.

Para el desarrollo de la interfaz gráfica que los usuarios utilizarán para clasificar los distintos comentarios según su carga ofensiva, se utilizará el *framework* *Angular* *7*, ya que éste es de código abierto, fácil implementación y altamente escalable (Angular, s.f.). Además, ofrece una tecnología capaz de mantener la aplicación en una sola página ofreciendo respuestas más rápidas a las interacciones del usuario. Para la integración con la red neuronal se utilizará *Django*, un *framework* web para *Python* compatible con *Angular 7*.

# Este proyecto se planea elaborar durante el período de diciembre 2018 - febrero 2019 y su finalidad será principalmente desarrollar una red neuronal cuya arquitectura retorna como resultado un número entre 0 y 1, el cual corresponde a la probabilidad de que un comentario sea considerado como ofensivo y será clasificado según el resultado como “Poco probable de ser ofensivo”, “relativamente neutro” y “Bastante probable de ser ofensivo”.

# Objetivo general y objetivos específicos

## General:

* Implementar un sistema de detección de la carga ofensiva de comentarios y mensajes en español en redes sociales.

## Específicos:

* Definir los criterios para evaluar la carga ofensiva de los comentarios.
* Calificar los datos extraídos de una red social seleccionada, a través de los criterios definidos.
* Diseñar el software basado en la Red Neuronal que detecta la carga ofensiva de comentarios en una red social
* Desarrollar el software bajo los *frameworks Django* y *Angular*

# Justificación

La labor que deben prestar los ingenieros de sistemas, por definición, consiste en solucionar problemas de carácter cotidiano para una población determinada. Esto se hace mediante una metodología de investigación tecnológica y práctica para generar soluciones a problemas del área tecnológica con el fin de aumentar las ganancias de las empresas. En este sentido, el principal motivo del desarrollo de la investigación consiste en solventar la actual demanda de herramientas para la clasificación de elementos ofensivos en la sección de comentarios de las publicaciones de las redes sociales. Actualmente, la red social *Instagram* funciona con una herramienta para comentarios en inglés llamada *Perspective*, la cual está siendo desarrollada por *Google* como una API privada. Sin embargo, se desea crear una herramienta para el español, que contribuirá con mantener un ambiente sano de debate en la comunidad hispanoamericana.

De acuerdo con lo expresado anteriormente, la capacidad para detectar y eliminar estos comentarios de las redes sociales, es inexistente en español, y según Cramer (2017), directora ejecutiva de la Sociedad Real para la Salud Pública, en la actualidad la red social más negativa es *Instagram*, según su estudio basado en personas jóvenes, se encuentra que el 91% de ellos han sido víctimas de comentarios ofensivos o *cyberbullying* y han reportado que no han tenido ninguna solución al respecto. Así mismo, el estudio muestra que en la actualidad 7 de cada 10 personas jóvenes son víctimas frecuentes de estos ataques. También, se afirma en la investigación que estas ocurrencias son un riesgo para la salud mental, la educación y las relaciones sociales, con un alto riesgo de que se mantenga en la adultez. (p.11). Por este motivo, muchas empresas están enfocadas en trabajar este problema. *Google* por su parte, desde el año 2017, ha puesto en desarrollo una API llamada *Perspective* (Jigsaw, 2017), enfocada en determinar casos de violencia en textos. Sin embargo, este proyecto está en fase de desarrollo y está limitado únicamente al idioma inglés.

Siguiendo el patrón de incremento de usuarios en las diferentes redes sociales, la demanda de calidad de servicio irá incrementando proporcionalmente con el pasar de los años. Un gran porcentaje de usuarios han experimentado ataques por medio de mensajes en las redes sociales, por lo que algunas empresas importantes están emprendiendo proyectos para combatir la agresividad. Sin embargo, se han visto muy limitados. Debido a esto, es necesaria una herramienta que clasifique comentarios ofensivos en español, contribuyendo con ofrecer un servicio de mayor calidad, logrando reducir los índices de agresión y que los usuarios permanezcan más tiempo en las aplicaciones.

# Metodología preliminar

El presente proyecto de investigación fue clasificado de acuerdo a su naturaleza y su alcance. Debido a su naturaleza, se deduce un enfoque cualitativo ya que Hernández, Fernández y Baptista (2014) exponen que las investigaciones cualitativas son aquellas cuyo objetivo principal es estudiar o evaluar las diferentes realidades de un tema en función de los actores que influyen en dicha realidad. En el caso propuesto, se estudiará y aplicarán conceptos para generar datos que ayuden a mejorar la experiencia de uso de los usuarios de redes sociales hispanoamericanos.

Por su alcance, la investigación es clasificada como descriptiva debido a que el estudio busca detectar la carga ofensiva que tienen los comentarios realizados en español en redes sociales. Esto permitirá aplicar diversas soluciones a la problemática mencionada anteriormente.

**Método de estudio**

        Este proyecto consiste en aplicar métodos conocidos en el área de inteligencia artificial para detectar patrones negativos o insultantes en los comentarios de las redes sociales para así potencialmente disminuir la problemática mencionada anteriormente. Debido a esto y dado el enfoque cualitativo que presenta, el proyecto se desenlaza en las siguientes etapas:

1. **Revisión de la literatura.**

La revisión bibliográfica estará enfocada en diferentes investigaciones en el área de redes neuronales, temas de desarrollo de software con inteligencia artificial y la documentación de las API de las redes sociales en cuestión y de los *frameworks* con los que se estará trabajando. Este proceso de revisión será elaborado a lo largo de toda la investigación y durante el desarrollo del software.

1. **Selección de población y muestra.**

Como se mencionó anteriormente, la investigación está centrada en la elaboración de un sistema de detección de la carga ofensiva de comentarios tóxicos que genere datos que ayuden a personas interesadas a estudiar el comportamiento de los usuarios de las redes sociales.

Nuestra población para el caso de estudio, comprende los comentarios en español de la red social seleccionada. Para definir la cantidad de datos a utilizar, Dixon, Thain y Wulczyn (2017) sugieren que la muestra sea representada por alrededor de 8000 comentarios verificados de los cuales al menos la mitad son calificados con anterioridad como ofensivos, esto con el fin de que la red neuronal tenga una base de comentarios ofensivos suficientemente grande para aprender a diferenciarlos de los no ofensivos.

1. **Recolección de datos**

El desarrollo del software se realizará mediante la metodología *Kanban*. Uno de los principios en los que se basa esta técnica es en estructurar la carga de actividades de forma que se mejora el flujo de trabajo mediante entregas funcionales continuas (Anderson, 2010).  Se utilizarán los principios básicos de esta metodología para garantizar un desarrollo de calidad en el tiempo limitado para la investigación.

La recolección de datos se realizará a través de las API de las redes sociales y de una aplicación desarrollada por los autores que extraiga una muestra significativa de comentarios en publicaciones de la red social seleccionada de usuarios de habla hispana. Posteriormente estos datos serán verificados por los investigadores, con el fin de filtrar aquellos comentarios no relevantes como pueden ser *emojis*, cadenas de caracteres aleatorios, mensajes conocidos como *spam* y aquellos que no estén en el idioma español. Adicionalmente se utilizará una interfaz gráfica para que los comentarios recolectados puedan ser calificados por usuarios comunes con el fin de establecer un aproximado de la probabilidad de que se califique como ofensivo. Dixon, Thain y Wulczyn (2017) sugieren que cada comentario puede tener hasta un máximo de 40 calificaciones de personas diferentes. Una vez clasificados todos los comentarios, se utilizarán las calificaciones para el cálculo de la probabilidad de que sea calificado como ofensivo dividiendo la cantidad de veces que fue calificado ofensivo entre la cantidad total de personas que lo calificaron.

1. **Procesamiento de los datos**

Utilizando la guía de Yao y Tan (2001) que explica una manera de trabajar con las redes neuronales, se expone que los datos se deben dividir en 3 grupos: 1. 70% de los datos, se utilizarán para el entrenamiento de la red. 2. 15% de los datos para validación del funcionamiento. 3. El 15% restante de los datos hacer una evaluación final de la red neuronal.

        La red neuronal, mediante el análisis de oraciones en lenguaje natural presentado en N-gramas, determinará la probabilidad de que un comentario sea calificado como ofensivo. Las probabilidades se clasificarán como “Poco probable de ser ofensivo”, “relativamente neutro” y “Bastante probable de ser ofensivo” dependiendo del resultado obtenido.

Una vez definida la arquitectura se procederá a desarrollar el modelo planteado para su posterior entrenamiento con los datos recolectados en la fase anterior. Finalmente se evaluará el funcionamiento de la red neuronal utilizando el bloque de datos designado para validación, con la finalidad de verificar que el modelo detecta correctamente datos externos al entrenamiento y comunicar los resultados mediante la interfaz gráfica.

1. **Comunicación de los resultados**

Luego de finalizado el proceso de desarrollo y entrenamiento, se procederá a verificar el funcionamiento de la red neuronal con el fin de determinar si logran detectar acertadamente la probabilidad de que los comentarios sean calificados como ofensivos. Esto se realizará a través de una interfaz gráfica que permitirá a los usuarios solicitar estadísticas de múltiples comentarios en las publicaciones de redes sociales. El software se encargará de procesar la data y enviarla a la red neuronal y una vez que ésta determine los resultados se le mostrarán al usuario para así concluir con una herramienta que cumpla con el objetivo principal de la investigación.

# Referencias

* Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow:* Large-Scale Machine Learning on Heterogeneous Distributed Systems. Recuperado de <http://download.tensorflow.org/paper/whitepaper2015.pdf>
* Anderson, D. (2010) The principle of the Kanban method. Recuperado de: https://web.archive.org/web/20140114161522/http://www.djaa.c om/principles-kanban-method
* Angular. (2018). Angular: *One framework. Mobile & Desktop.* Recuperado de: <https://angular.io/docs>
* Beck, K (1999) Extreme Programming Explained. Recuperado de: <http://ptgmedia.pearsoncmg.com/images/9780321278654/samp> lepages/9780321278654.pdf
* [Bruguier](https://www.semanticscholar.org/author/Antoine-Bruguier/2514424), A., [Damavandi](https://www.semanticscholar.org/author/Babak-Damavandi/3057557), B., [Kumar](https://www.semanticscholar.org/author/Shankar-Kumar/9567965), S., [Shazeer](https://www.semanticscholar.org/author/Noam-Shazeer/1846258), N. (2016). *NN-grams*: Unifying neural network and n-gram language models for speech recognition. San Francisco, USA. Recuperado de https://pdfs.semanticscholar.org/5b7a/b3fe5d4684289280f52a8 2a5709a92ab2caf.pdf?\_ga=2.184404644.320230059.15272010 06-591438879.1527201006
* Cramer, S. (2017) #*StatusOfMind Social media and young people’s mental health and wellbeing*. Recuperado de:

https://www.rsph.org.uk/uploads/assets/uploaded/62be270a- a55f-4719-ad668c2ec7a74c2a.pdf

* Dixon, L., Thain, N., Wulczyn, E. (2017). *Ex Machina: Personal attacks seen at scale*. Ithaca, New York: Cornell University. Recuperado de https://arxiv.org/abs/1610.08914
* García, F. (2018). *La investigación tecnológica*: Investigar, indagar e innovar en ingenierías y ciencias sociales. (2da edición). México: Limusa.
* Gers, F., Schmidhuber, J., Cummins, F. (1999). Learning to Forget: *Continual Prediction with LSTM.* Lugano, Suiza. Recuperado de: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1660E> CB8AAEFCE2517AB996099A3F8C1?doi=10.1.1.55.5709&rep= rep1&type=pdf
* Gers, F., Schmidhuber, J. (sf.). LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages. Manno, Suiza. Recuperado de: <ftp://ftp.idsia.ch/pub/juergen/L-IEEE.pdf>
* Hernández, R., Fernández, C. & Baptista, P. (2014). Metodología de la investigación. 6ta ed. México: McGraw-Hill
* Jigsaw (2017). Perspective API. Recuperado de: https://www.perspectiveapi.com/#/
* Keras (2015). *Keras*: The Python Deep Learning library. Recuperado de <https://keras.io/>
* Kohonen, T. (1989). *Self-organization and associative memory*. Springer-Verlag, New York. Recuperado de

https://ia801904.us.archive.org/28/items/SelfOrganizationAndAs sociativeMemory/Self%20Organization%20and%20Associative %20Memory\_text.pdf

* Python (2018) The Python Tutorial. Recuperado de: <https://docs.python.org/3/tutorial/index.html>
* Real Academia Española (2001). Diccionario de la lengua española (22da edición). Recuperado de:

http://lema.rae.es/drae/?val=lenguaje

* Rebollo, M. (2018, septiembre 29). Quora: *¿Cuáles son los mejores lenguajes de programación para machine learning?* [comentario en un foro en linea]. Recuperado de: https://es.quora.com/Cu%C3%A1les-son-los-mejores- lenguajes-de-programaci%C3%B3n-para-machine-learning
* Stamateas, B. (2008). *Gente toxica*: Las personas que nos complican la vida y cómo evitar que sigan haciéndolo. (1era edición). Barcelona, España: Ediciones B.

https://www.fisterra.com/mbe/investiga/cuanti\_cuali/ cuanti\_cuali.asp

* Yao, J. & Tan, C. (2001). *Guidelines for financial forecasting with neural networks. Proc*. International Conference on Neural Information Processing, Shanghai, China, pp 757-761.

# Cronograma de trabajo semanal

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Semana | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 9 | 10 | 11 | 12 |
| Actividad |  | 19/11/18 al 25/11/18 al | 26/11/18 al 02/12/18 | 03/12/18 al 09/12/18 | 10/12/18 al 16/12/18 | 17/12/18 al 23/12/18 | 24/12/18 al 30/12/18 | 31/12/18 al 06/01/19 | 07/01/19 al 13/01/19 | 14/01/19 al 20/01/19 | 21/01/19 al 27/01/19 | 28/01/19 al 03/02/19 |
| Diseño de la arquitectura del software e interfaces graficas | |  |  |  |  |  |  |  |  |  |  |  |
| Recolección de comentarios de la red social seleccionada | |  |  |  |  |  |  |  |  |  |  |  |
| Calificación de los comentarios | |  |  |  |  |  |  |  |  |  |  |  |
| Desarrollo del modelo de red neuronal y la arquitectura del software. | |  |  |  |  |  |  |  |  |  |  |  |
| Entrenamiento del modelo. | |  |  |  |  |  |  |  |  |  |  |  |
| Validación y pruebas del modelo. | |  |  |  |  |  |  |  |  |  |  |  |
| Integración de la interfaz gráfica para la comunicación de resultados. | |  |  |  |  |  |  |  |  |  |  |  |
| Redacción del tomo | |  |  |  |  |  |  |  |  |  |  |  |

# 

# Planificación del trimestre a futuro

Maximiliano Casale:

Trimestre 1819-2

* Defensa de Tesis

Andrés Medina:

Trimestre 1819-2

* Defensa de tesis