# Fiero: Asistente virtual para la captación de insultos

## *Fiero: A virtual assistant for collecting insults*

**Beatriz Botella Gil**[1], **Flor Miriam Plaza del Arco**[2], **Ana Belén Parras Portillo**[2], **Yoan Gutiérrez**[1]
[1]Instituto Universitario de Investigación Informática, Universidad de Alicante, España
{beatriz.botella, ygutierrez}@dlsi.ua.es
[2]Departamento de Informática, CEATIC, Universidad de Jaén, España
{fmplaza, abparras}@ujaen.es

**Resumen:** Fiero es un asistente virtual orientado a la recopilación de insultos, expresiones vulgares, comentarios ofensivos o cualquier forma de lenguaje no aceptable a través de la aplicación de mensajería Telegram. A través de esta aplicación, los usuarios pueden tener una conversación real con Fiero donde se incita a que lo insultemos de forma humorística y sarcástica. Se ha puesto a disposición de la población obteniendo una gran variedad de improperios utilizados en español. La recopilación de estos insultos será fundamental para la creación de recursos lingüísticos que podrán ser posteriormente utilizados para ser integrados en sistemas computacionales con el fin de identificar comportamientos inapropiados en la Web como, por ejemplo, el ciberacoso o el discurso del odio en sus diferentes formas.
**Palabras clave:** Aistente virtual, Bot, Lenguaje ofensivo, Telegram.

**Abstract:** Fiero is a virtual assistant aimed at collecting insults, vulgar expressions, offensive comments or any form of unacceptable language across the messaging application Telegram. Users can chat with Fiero where it encourages them to insult in a humorous and sarcastic way. It has been made available to the Spanish population obtaining a wide variety of expletives used in Spanish. The collection of these in-sults will be essential for the creation of linguistic resources in Spanish. In addition, their integration in systems based on Human Language Technologies could help to identify social problems present on the Web such as cyberbullying or hate speech in its different manifestations.
**Keywords:** Virtual Assistant, Bot, Offensive Language, Telegram.



## *1 Introducción y Motivación*

Las TICS (Tecnologías de la Información y la Comunicación) se han asentado en nuestra sociedad cambiando la forma de comunicarse entre las personas. Es una realidad que la era digital está aportando grandes beneficios a la sociedad, pero también el aumento de las interacciones sociales digitales y el anonimato, ha promovido la presencia de conductas y mensajes violentos en las Web.

En referente al ciberacoso, según la encuesta (Sanjuán et al., 2019) realizada por Save the children , el 40 % de los jóvenes han sufrido este tipo de acoso y la recepción de estos mensajes violentos empezó a los 8 y 9 años. Esta edad temprana se debe al acceso en aumento que tienen los menores en España, en concreto, según el INE, la población entre 16 y 24 años usa las TICS en un 99,8 % como refleja en las últimas estadísticas realizada en 2020 (INE, 2020).

Al mismo tiempo, este problema también implica a los gobiernos y las plataformas digitales. Por ello, para combatir la difusión del lenguaje ofensivo en la Web, continuamente se están desarrollando leyes y políticas de lucha contra la incitación al odio. Desde 2013, el Consejo Europeo ha promovido el movimiento "No Hate Speech", con el objetivo de movilizar a los jóvenes para combatir el discurso de odio y defender los derechos humanos en Internet. En 2016, la Comisión Europea llegó a un acuerdo con Facebook, Microsoft, Twitter y YouTube para crear un "Código de conducta sobre la lucha contra el discurso de odio ilegal en Internet"[1]. Según un

[1]https://cutt.ly/Hj5EsAh

informe español de 2019 sobre la evolución de los delitos de odio en España [2], las amenazas, los insultos y la discriminación se contabilizan como los actos delictivos más repetidos, siendo Internet (54,9 %) y las redes sociales (17,2 %) los medios más utilizados para cometer estas acciones. Este informe llevó al parlamento español a aprobar en 2020 un proyecto de ley para evitar la propagación del odio en la red[3].

El Procesamiento del Lenguaje Natural (PLN) desempeña un papel fundamental en la detección de este tipo de contenido en la Web ya que permite desarrollar sistemas computacionales que ayuden a procesar e interpretar el lenguaje humano. Para entrenar estos sistemas, es necesario disponer de recursos específicos para la tarea objetivo, en este caso, la identificación del lenguaje ofensivo (Plaza-del Arco et al., 2019), (Plaza-Del-Arco et al., 2020a). En los últimos años, la comunidad científica del PLN ha invertido considerables esfuerzos en la creación de este tipo de recursos (Zampieri et al., 2019), (Wiegand y Siegel, 2018), (Plaza del Arco et al., 2020b). Sin embargo, la mayoría están disponibles para el inglés, lo que conlleva la necesidad de crear recursos lingüísticos en otros idiomas cuya presencia es notable en la Web, como el español.

En este artículo se presenta Fiero, un asistente virtual que mantiene una conversación con el usuario animándolo a expresar improperios a través de Telegram. Esta aplicación es popularmente conocida en el ambiente propio del uso de herramientas digitales por parte de la población. El diálogo recopilado servirá para generar recursos lingüísticos que se puedan usar en sistemas automáticos de inteligencia artificial para combatir problemas sociales como el ciberacoso o la propagación del discurso del odio.

## *2 Fiero*

### 2.1 Características

El asistente virtual Fiero se ha desarrollado teniendo en cuenta diferentes características en base a su utilización por parte de los usuarios:

- **Recopilación de datos demográficos del usuario**. Se recopilan dos variables, la edad (mayor o menor de 18 años) y el sexo (hombre o mujer) al iniciar la aplicación, con el objetivo de identificar los comentarios utilizados para la expresión del lenguaje ofensivo por diferentes sectores de la población.
- **Anonimización**. Al tratarse de un programa diseñado para comunicarse con personas y recopilar un gran volumen de datos, es necesario asegurar a los usuarios su privacidad para un uso seguro y confiable de la aplicación. Por ello, Fiero solo recopila su género y edad, garantizando la preservación de la privacidad y el derecho a la protección de datos personales en todo momento.
- **Diálogo**. Fiero establece los diálogos apoyándose en una lista de preguntas-respuestas en un único contexto de *DialogFLow*. Se trata de una herramienta de creación de *chatbots* capaz de entender el lenguaje natural y que provee infraestructura para recrear conversaciones y construir diálogos con el fin de interactuar con el usuario de manera fluida (Sabharwal y Agrawal, 2020). Los contextos de Dialogflow son similares al contexto del lenguaje natural. Las conversación entre el usuario y Fiero consiste en los siguientes pasos:
  - El usuario escribe una entrada: el mensaje.
  - El agente (o módulo de compresión de lenguaje natural) extrae cada uno de los parámetros de dicha entrada. En este paso es donde se le solicita al usuario una serie de preguntas del tipo demográfico y a continuación se le anima de forma humorística y sarcástica para que escriba insultos.
  - El agente devuelve la respuesta (previamente programada) gracias a *DialogFlow* que se corresponde con la entrada del usuario.

### 2.2 Arquitectura

El componente principal en la arquitectura de Fiero es un componente que actúa como controlador, u orquestador, de procesos mediante el cual el flujo de información entre el usuario y la parte servidora de la aplicación. A través de este componente:

[2] `https://cutt.ly/ej5EgU7`
[3] `https://cutt.ly/5j5Ejum`

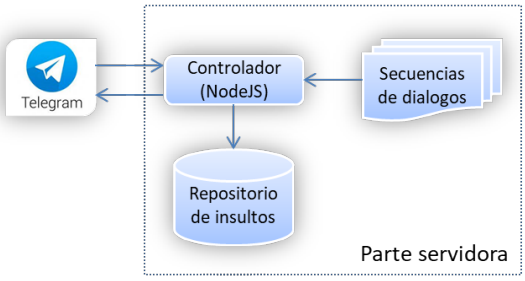


Figura 1: Arquitectura de Fiero

- la parte servidora interactúa con la plataforma de Telegram.
- se construyen los diálogos. Para esto contamos con un diccionario de secuencia de diálogos afines con distintos tipos de preguntas. En este proceso técnicas simples de PLN se aplican para emparejar insultos de los usuarios con determinadas respuesta del sistema que utilizan jergas populares, logrando motivar al usuario a insultar al bot.
- se almacena información proveniente del usuario en el repositorio, los datos demográficos y los comentarios provenientes del diálogo.

Los módulos de la parte servidora utilizan NodeJS[4] como tecnología de programación.

### 2.3 Interfaz de Usuario

El usuario activamente puede mantener una conversación con Fiero a través de la interfaz intuitiva que ofrece Telegram. En la Figura 2.3 se muestra la simulación de un diálogo real con Fiero. En primer lugar, se solicita la edad y el sexo, una vez proporcionados estos datos, el asistente virtual ofrece el reto al usuario de enfadarlo a través de improperios.

## *3 Estadísticas de Uso*

Fiero está continuamente activo para interactuar con los usuarios a través de la aplicación de mensajería de Telegram. Se puso a disposición de los usuarios en el mes de julio de 2019 y hasta febrero de 2021 se han recopilado una serie de estadísticas que recogen la utilización de Fiero por parte de la población. La Tabla 1 muestra el número de comentarios obtenidos en dicho período de tiempo. Cabe destacar que durante el año 2020 hubo una interacción notablemente mayor, obteniendo

[4]https://nodejs.org/

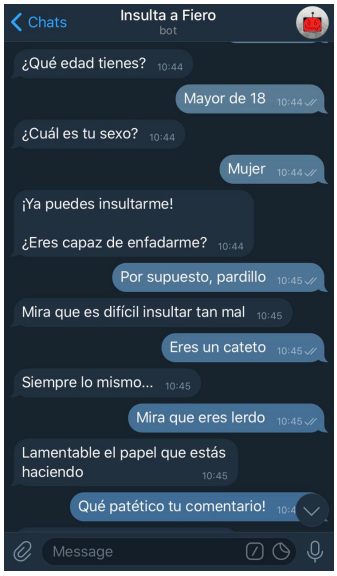


Figura 2: Interfaz de usuario de Fiero

un total de 150.754 comentarios debido a la gran difusión y repercusión de Fiero en los medios de comunicación de radio, prensa y televisión nacional. Finalmente, se obtiene un total de 164.467 comentarios en el periodo de tiempo mencionado (2019-2021).

Con respecto a las variables demográficas, se han recopilado el número de comentarios obtenidos por mujeres y hombres menores y mayores de 18 años (Tabla 2). Cabe destacar que la población masculina mayor de 18 años ha realizado una mayor interacción con Fiero recopilando un total de 95.513 comentarios. La población más joven (<18) participa en menor medida, obteniendo un total de 17.037 comentarios en comparación con 147.430 comentarios obtenidos por el sector mayor de la población.

| Año | #Comentarios |
|---|---|
| 2019 | 102 |
| 2020 | 150.754 |
| 2021 | 13.611 |
| Total | 164.467 |

Tabla 1: Número de comentarios obtenidos por año en Fiero

Con la intención de efectuar un análisis del lenguaje natural, se ha realizado un estudio para obtener el total de comentarios correspondientes a unigramas, bigramas, trigramas (donde ya el usuario ha utilizado expresiones), e incluso n-gramas, simulando una posible conversación con el bot. Dichos datos

| Sexo | Edad | #Comentarios |
|---|---|---|
| Mujer | >18 | 51.917 |
| | <18 | 5.922 |
| Hombre | >18 | 95.513 |
| | <18 | 11.115 |
| Total | | 164.467 |

Tabla 2: Número de comentarios obtenidos según sexo y edad en Fiero

se exponen en la Tabla 3. La mayor proporción (55,33 %) corresponde a unigramas y la mayor parte de ellos se refieren a insultos a Fiero, seguido de los n-gramas (25,65 %) donde se observa un uso más rico del lenguaje al entablar una conversación más real.

| N-grama | #Comentarios |
|---|---|
| Unigramas | 91.005 |
| Bigramas | 16.613 |
| Trigramas | 14.649 |
| N-gramas | 42.200 |
| Total | 164.467 |

Tabla 3: Número de comentarios distribuidos en ngramas en Fiero

## *4 Conclusiones y Trabajo futuro*

En este artículo se presenta Fiero, un asistente virtual accesible desde Telegram que simula una conversación real con el usuario para recopilar insultos, expresiones vulgares o cualquier forma de comentario no aceptable. Esta herramienta se ha puesto a disposición de la población obteniendo una gran variedad de comentarios que recogen los improperios más utilizado en el registro español. El principal objetivo marcado como trabajo futuro es recopilar estos improperios para la creación de recursos lingüísticos en esta lengua. Su integración en sistemas basados en TLH que permitirá el desarrollo de sistemas automáticos que ayudarán a identificar problemas sociales presentes hoy en día en la Web como el discurso del odio o el ciberacoso.

## *Agradecimientos*

Esta investigación ha sido parcialmente financiada por la Universidad de Alicante, la Universidad de Jaén, el Ministerio de Ciencia, Innovación y Universidades (Beca FPI-PRE2019-089310) y el Ministerio de Economía y Competitividad del Gobierno de España, a través de los proyectos LIVING-LANG (RTI2018-094653-B-C21, RTI2018-094653-B-C22) , SIIA (PROMETEU/2018/089), e INTEGER (RTI2018-094649-B-I00).

## *Bibliografía*

INE. 2020. Encuesta sobre equipamiento y uso de tecnologías de información y comunicación en los hogares.

Plaza-del Arco, F. M., M. D. Molina-González, M. T. Martín-Valdivia, y L. A. U. Lopez. 2019. SINAI at semeval-2019 task 6: Incorporating lexicon knowledge into svm learning to identify and categorize offensive language in social media. páginas 735–738.

Plaza-Del-Arco, F.-M., M. D. Molina-González, L. A. Ureña-López, y M. T. Martín-Valdivia. 2020a. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*.

Plaza del Arco, F. M., C. Strapparava, L. A. Urena Lopez, y M. Martin. 2020b. EmoEvent: A multilingual emotion corpus based on different events. En *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association.

Sabharwal, N. y A. Agrawal. 2020. Introduction to Google Dialogflow. En *Cognitive Virtual Assistants Using Google Dialogflow*. Apress, páginas 13–54.

Sanjuán, C., A. S. Campo, C. del Moral, M. Pereda, B. Irene, A. Montiel, J. Greco, N. M. Hombrado, P. Cabrera, y Óscar Naranjo. 2019. Violencia viral.

Wiegand, M. y M. Siegel. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. En *Proceedings of KONVENS 2018*.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, y R. Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). En *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.