COMPILACIÓN Y ETIQUETADO DE CORPUS PARA EL ANÁLISIS DE LA VIOLENCIA LINGÜÍSTICA EN TWITTER: PROBLEMAS Y SOLUCIONES

COMPILATION AND TAGGING OF CORPORA FOR THE ANALYSIS OF LINGUISTIC VIOLENCE ON TWITTER: PROBLEMS AND SOLUTIONS

Susana María Campillo Muñoz
M. Dolores Jiménez López
Universitat Rovira i Virgili

RESUMEN
El análisis de la violencia verbal cada vez tiene más presencia tanto en los estudios lingüísticos como en los computacionales. Sin embargo, ambas disciplinas todavía encuentran algunos problemas relacionados, sobre todo, con la terminología empleada y las categorías que se consideran. En el ámbito computacional, en concreto, surgen problemas de análisis, tanto en lo referente al etiquetado como a los rasgos lingüísticos que se contemplan. Con el objetivo de identificar las dificultades que puedan surgir y proponer soluciones, hemos realizado una simulación de etiquetado en un corpus de muestra de 100 tuits de violencia verbal en español. Se ha etiquetado cada tuit como violento o no violento sin guía de anotación. Los resultados confirman la falta de consenso en la comprensión del concepto de violencia verbal y problemas en el análisis de frecuencias relacionados con los *hashtags* y los emoticonos. Tras analizar los resultados de esta simulación, presentamos algunas soluciones: utilizar una guía de anotación con definiciones concretas en la fase de etiquetado y una lista de atributos en diferentes niveles lingüísticos en el análisis computacional.

PALABRAS CLAVE: lingüística de corpus, violencia lingüística, análisis computacional

ABSTRACT
The interest in the study of verbal violence is increasing in linguistics and computational branches. However, investigators try to face some issues, such as conceptual definitions, and the categories included in the analysis. In computational analysis specially, there are problems with the annotation task and the linguistic features extracted. In order to detect the problems in verbal violence corpora analysis and to propose some solutions, we simulate a tagging task with a sample corpus of 100 tweets between three annotators. They must tag every tweet as violent or non-violent. Our results confirm these differences in the comprehension of verbal violence and some problems related to hashtags and emojis in the computational analysis. Then, we propose some solutions related to the annotation task and the computational analysis. With the aim of getting a common concept of verbal violence in the annotation task, we need to use an annotation scheme. Also, it

is necessary to create a list of different linguistic features, from emojis to situational attributes, for improving the computational analysis. To sum up, linguistics and computation need to work together so that we could achieve best results in the analysis of verbal violence.

KEYWORDS: corpus linguistics, linguistic violence, computational analysis.

1. INTRODUCCIÓN

En la actualidad, la violencia verbal es un fenómeno presente en muchos ámbitos de nuestra sociedad. Esto, junto con el auge de las redes sociales y la facilidad de publicar mensajes al instante, ha provocado que distintas disciplinas se interesen por el análisis de la violencia verbal. Así, las empresas que gestionan las redes sociales se han preocupado por crear normas de uso en las que se concreta qué tipo de mensajes están prohibidos según unos criterios generalmente consensuados. El objetivo es que los algoritmos sean capaces de detectar estos mensajes para poder mediar las interacciones en la red.

Desde la lingüística, por un lado, se estudia el fenómeno bajo los términos *descortesía*, *agresividad verbal* o *violencia verbal*. Desde el ámbito de la computación, por otro lado, es común el término *hate speech* o *discurso de odio* que, aunque hace referencia a un tipo de violencia concreto, permite abordar el fenómeno de la violencia verbal desde el procesamiento del lenguaje natural. Anualmente se realizan *workshops* en todo el mundo (*Workshop on Abusive Language Online*, *Workshop on NLP for Internet Freedom*, *Workshop on Trolling, Aggression and Cyberbullying*, entre otros) en los que se presentan diferentes algoritmos que pretenden detectar y analizar automáticamente los mensajes que contienen *hate speech*. Los resultados que se obtienen ofrecen una visión amplia de los problemas a los que se enfrentan. Así, ambas disciplinas trabajan en paralelo con un objetivo común: el análisis de la violencia verbal.

No obstante, ambas disciplinas se encuentran con problemas terminológicos y analíticos parecidos. Uno de los retos es el análisis de mensajes implícitos, como aquellos que contienen ironía o metáforas. Aunque desde la lingüística se han incluido estos tipos de mensajes en el estudio de la violencia verbal, en la computación es ahora cuando se empiezan a analizar. La presencia de falsos negativos, esto es, de aquellos mensajes que el algoritmo no es capaz de detectar, ha provocado que el interés por el estudio de lo implícito sea cada vez mayor. El reto reside en conseguir que el algoritmo sea capaz de comprender, del mismo modo que hacemos los hablantes, los mensajes violentos en sus distintas materializaciones.

Es en este punto en el que creemos que sería lógico que ambas disciplinas se conectaran: el análisis computacional permite a la lingüística trabajar con gran cantidad de datos y las teorías lingüísticas facilitan a la computación un marco teórico sólido relacionado con la violencia verbal y la comunicación implícita. Puesto que en este momento se encuentran con problemas similares, sería eficaz que trabajaran en conjunto para resolver dichas dificultades.

Este artículo se estructura de la siguiente manera: en el apartado 2, realizamos una breve explicación de los estudios lingüísticos y computacionales sobre violencia verbal y *hate speech* para mostrar una panorámica general de las investigaciones actuales; en el apartado 3, explicamos la metodología seguida para compilar, etiquetar y analizar el corpus de nuestra prueba piloto; en el apartado 4, exponemos los resultados de dicha simulación, incidiendo en las dificultades encontradas; en el

apartado 5, comparamos las propuestas de las últimas investigaciones sobre etiquetado y análisis de corpus de *hate speech*, y proponemos algunas soluciones para solventar los problemas hallados en nuestra simulación; en el apartado 6, explicamos nuestra propuesta: la guía de anotación y la lista de rasgos lingüísticos y, finalmente, en el apartado 7, destacamos los aspectos relevantes de nuestra simulación y propuesta.

2. ESTADO DE LA CUESTIÓN

2.1. Estudios lingüísticos sobre violencia verbal

La violencia verbal ha sido tratada tradicionalmente desde la lingüística bajo el término *descortesía*. Esta se nutrió de la Teoría de la Cortesía (Brown y Levinson 1987) y, por tanto, partió de una perspectiva anglocéntrica (Bravo, 2004). Culpeper (1996) tomó el concepto de *face* del que partieron Brown y Levinson: la imagen está formada por 'la imagen positiva' (la imagen social) y 'la imagen negativa' (la libertad de acción), que hay que salvaguardar. A partir de esta distinción, estableció diferentes estrategias de descortesía según la imagen a la que atacan y el nivel de explicitud: descortesía directa (*Bald-on-record impoliteness*) y, dentro de esta, descortesía positiva (*Positive impoliteness*) y descortesía negativa (*Negative impoliteness*), falsa cortesía (*Sarcasm or mock politeness*), descortesía encubierta (*Off-record impoliteness*) y acortesía (*Withhold politeness*). Esta clasificación fue el punto de partida para los estudios posteriores.

Tras la teoría de la descortesía (Culpeper, 1996), fueron varios las investigaciones que surgieron en relación con estos enunciados: se centraban en varios géneros (interrogatorios, correos electrónicos, entrevistas) y en diferentes formatos de corpus (escrito, oral o virtual). Los estudios que desencadenó dicha teoría se centraron, en general, en aplicar la propuesta de Culpeper (1996) al análisis discursivo de corpus de diferente procedencia, traduciendo, así, las categorías que contemplaba dicha teoría y realizando un análisis manual de corpus muy concretos. Estos corpus eran de diferentes lenguas, hecho que conllevó la aplicación de una clasificación propia de determinado contexto sociocultural a otro distinto, por lo que era probable que no se correspondiera con dicha cultura.

Este interés que suscitó la descortesía hizo que la teoría inicial evolucionara, pues los resultados que se obtenían no acababan de ser generalizables y aplicables a otros contextos. En ese punto fue de especial importancia la evolución de los estudios sociopragmáticos sobre la cortesía y la descortesía, pues cuestionaron las concepciones base que tomaban las teorías de Brown y Levinson, y de Culpeper.

Tras la teoría de la descortesía, estudios como los de Terkourafi (2008) y los de Bravo (1999) explicitaron la visión occidental de dichas teorías, pues tomaban como base el concepto de imagen o *face* (Goffman, 1955) anglosajón, con dos categorías que completaban esta imagen ('imagen negativa' e 'imagen positiva') y la delimitaban claramente. En el mundo hispánico, Bravo (1999, 2004) propuso una alternativa al concepto de *imagen* del que partía Culpeper (1996): la *autonomía* y la *afiliación*. Así, partiendo de que el acto comunicativo es una actividad de imagen, Bravo (1999) defendió la *imagen* como un concepto vacío que cada cultura rellena y completa de forma distinta, y que permite explicitar el conocimiento compartido sobre los efectos sociales de determinadas acciones. Esta visión sociopragmática, junto con las actividades de imagen más comunes en la cultura española (resaltar las buenas

cualidades ajenas, mostrar reciprocidad y generosidad, estar cohesionados con el grupo social al que se pertenece, etc.), permiten analizar el concepto de *descortesía* considerando el contexto sociocultural e incluyéndolo como elemento fundamental en su estudio.

Cabe destacar, además, la confusión entre el término teórico *descortesía* y el concepto más subjetivo que se tiene de esta (Culpeper, 2017). Se entremezclan las comprensiones subjetivas, basadas en la experiencia individual de cada hablante, con las concepciones objetivas, propias del estudio del fenómeno desde el prisma científico. A ello cabe sumar las diferencias culturales en torno a la sanción social y permisibilidad de diferentes estrategias de descortesía.

En el contexto español, la descortesía se concibe en general como *mala educación* o *grosería* (Bou-Franch, 2021), hecho por el que puede no incluir categorías de mayor grado como el rechazo. Es por ello por lo que en nuestra investigación empleamos el término *violencia verbal* como genérico. Las definiciones de violencia comparten elementos clave tales como el poder de dañar física o emocionalmente; la sanción social, moral y legal, y la intención o percepción de esta (Henry y Milovanovic, 2000; Scheper-Hughes y Bourgois, 2004). Así, a partir de varias definiciones de violencia y de *descortesía* de distintos ámbitos, concebimos la *violencia verbal* como cualquier acto verbal que daña la imagen de la persona receptora o referida en mayor o menor grado y que es percibido como intencionado. No especificamos en la definición de violencia verbal las categorías que incluye.

2.2. Estudios computacionales sobre *hate speech*

En paralelo con la lingüística, el análisis computacional se ha centrado en el llamado *hate speech* en las redes sociales. Debido al auge en el uso de estas y sus características (posible anonimato, distancia físico-temporal, rapidez) los usuarios han ido realizando contribuciones cada vez más violentas en plataformas como Twitter. Varias de estas plataformas y organizaciones definieron, a causa del incremento de la violencia verbal, normas de uso. Sin embargo, tanto las definiciones de dichas normas como las de los estudios computacionales son dispares y no están consensuadas (Fortuna. 2017). Esto, junto con el interés hacia la detección automática para regular estas publicaciones, hicieron que el ámbito computacional se interesara por el análisis de corpus de las redes sociales, en especial, Twitter.

La diversidad de definiciones del término *hate speech* ha sido destacada por autores como Fortuna (2017, 2018) y Poletto y otros (2020). La mayoría son en inglés y aspectos como atacar, dirigirse a una persona o grupo concreto e incitar a la violencia son compartidos en estas definiciones (Fortuna, 2017). Se incluyen también varias categorías con límites difusos entre ellas, como *abusive language*, *discrimination*, *toxic language* o *profanity*.

Esta variedad de definiciones y de categorías que se incluyen en el término genérico *hate speech* provoca que los estudios y análisis que se realizan no obtengan, en general, resultados generalizables, propios y característicos del *hate speech*, ni extrapolables, es decir, aplicables a otros corpus de tipología y procedencia distinta.

A pesar de que guarda diferencias con la *descortesía* (está dirigido a grupos en concreto llamados *categorías protegidas*, esto es, que parten de diferencias de género, orientación sexual, raza, procedencia y religión), el término *hate speech* se incluye dentro de la violencia verbal que consideramos en nuestra investigación. Pensamos que el acto es el mismo y que difieren los elementos situacionales (destinatario, rasgo que se destaca, temática). Tomamos, pues, la *violencia verbal* un término genérico y

difuso, con grados, no concreto. Esta gradación favorece la inclusión de categorías o clases consideradas como poco violentas y permite obtener una visión amplia del fenómeno y de sus características esenciales.

Con el objetivo de compartir y evaluar diferentes propuestas para el análisis automático del *hate speech*, anualmente se realizan diferentes *workshops* dirigidos al análisis computacional de corpus etiquetados para construir algoritmos con mayor porcentaje de acierto. En estos, grupos de investigadores proponen la combinación de algoritmos (tradicionales, como *Logistic Regression, Support Vector Machine* o *Random Forest*, o de redes neuronales, como *Convolutional Neural Network o Recurrent Neural Network*) y consideran distintos rasgos en el análisis (generalmente, estadísticos, aunque también lingüísticos).

De entre los rasgos incluidos en el análisis, aquellos que consideran rasgos lingüísticos (tanto insultos como presencia de estructuras morfosintácticas concretas) emplean rasgos generales, esto es, no prototípicos del *hate speech* y comunes en las técnicas de *text mining* (Fortuna, 2017). Por ello, es difícil no confundir rasgos lingüísticos propios de mensajes en determinada red social con rasgos lingüísticos propios del fenómeno de estudio.

Asimismo, todavía se obtienen, en los resultados, falsos negativos (mensajes que contienen *hate speech* y que no se detectan). Se destaca la dificultad de analizar mensajes implícitos, cuyo odio no está explícito, no está presente lingüísticamente en el enunciado. Recientemente, algunas investigaciones proponen clasificaciones en términos de explicitud e implicitud [abuso directo, generalizado, explícito o implícito (Waseem y otros, 2017)], así como de categorías de implícito [ironía y metáfora (Plaza-del-Arco y otros, 2020)], comparaciones y construcciones eufemísticas (Wiegand y otros, 2021)].

Sin embargo, sigue habiendo problemas en la detección de los mensajes implícitos. De entre las posibles causas, destacamos las relacionadas con:

- el etiquetado de corpus: no se especifica el método de etiquetado o se realiza sin una guía que defina los términos claramente;
- las definiciones: se especifican categorías que excluyen las no mencionadas, son ambiguas o apelan a aspectos subjetivos como la actitud (Assimakopoulos y otros, 2020), y
- las clasificaciones de las que se parte: mezcla de categorías de *hate speech* con niveles de explicitud, tales como *direct abuse, explicit abuse, generalized abuse*, (Wasseem y otros, 2017) o definiciones poco claras de la distinción explícito-implícito.

Para conseguir un análisis de mensajes implícitos eficaz, es necesario cuestionarse estas definiciones y clasificaciones.

3. METODOLOGÍA

Con el objetivo de observar los problemas a los que nos enfrentamos en el etiquetado de un corpus de violencia verbal, se ha realizado una prueba piloto. En este primer etiquetado, se ha utilizado un corpus compilado automáticamente atendiendo a varios criterios, tales como variedad temática y la información situacional necesaria en la red social.

El corpus utilizado se ha extraído de Twitter mediante la consola de Python utilizando la extensión Twint. Esta herramienta ha permitido filtrar mensajes por *hashtag* o mención, por idioma y por ubicación. En un rango de 10 días (del 1 al 10

de mayo de 2020), se han seleccionado de los *trending topic* de cada día aquellos relacionados con temas políticos, sociales, de fútbol o de programas mediáticos. Se han seleccionado estos temas porque creemos que son susceptibles de dar lugar a violencia verbal. De cada uno de los *trending topic* escogidos para cada día, se han extraído 100 mensajes. Así, el corpus posee 10.000 mensajes de Twitter. De cada mensaje, se ha extraído el nombre de la cuenta y del usuario, la fecha y la hora, el mensaje, los *retweets,* los *replies* y los *megusta*, las menciones y los *hashtags*. Toda esta información se ha recogido en una hoja de cálculo con distintas columnas para cada tipo de información.

Para nuestra prueba piloto, del total del corpus, se han extraído 100 mensajes de diferentes temáticas para hacer un primer etiquetado. Las menciones o *hashtag* seleccionados son @IreneMontero, #SABADODELUXE, #NazisyCorruptos, #Subnormales y #AlviseFollaardillas. De cada uno, se han seleccionado 20 ejemplos equitativamente, tal y como muestra la Tabla 1:

Tabla 1. *Trending topic*, temáticas correspondientes y tuits recopilados de cada una

Trending topic	Temática	Tuits
@IreneMontero	Política	20
#SABADODELUXE	Televisión	20
#NazisyCorruptos	Política	20
#Subnormales	Covid	20
#AlviseFollaardillas	Política	20

El etiquetado lo han realizado 3 anotadores, lingüistas, usuarios de Twitter y de un rango de edad de entre 25 y 35 años. Este primer etiquetado se ha llevado a cabo sin guía de anotación y se ha marcado cada tuit como violento (1) o no violento (0). El grado de acuerdo obtenido es de 0,53 utilizando la medida Krippendorff's Alpha.

Por último, el análisis computacional se ha realizado mediante Python con el paquete NLTK. Se han eliminado algunas de las *stopwords* (preposiciones, determinantes, puntos y comas) y se ha contemplado en el análisis los ítems más frecuentes, los n-gramas y las concordancias. Se muestran en la Tabla 2 las particularidades de cada análisis.

Tabla 2. Tipos de análisis realizado, ítems considerados y finalidad de cada tipo análisis

Tipo de análisis	Ítems	Finalidad
Ítems más frecuentes	*Tokens* delimitados por espacios: palabras, emoticonos y signos de puntuación.	Extraer las unidades más frecuentes con implicaciones lingüísticas.
N-gramas	Palabras y estructuras repetidas en un orden determinado. Conjuntos de tres (trigramas).	Obtener las cadenas de estructuras similares.
Concordancias	Unidades y cotexto en el que aparecen.	Observar la posición en la que aparece determinado ítem.

Estos tipos de análisis nos permiten observar recurrencias en el corpus que pueden tener implicaciones teóricas y analíticas.

4. RESULTADOS DE LA SIMULACIÓN

Tras el etiquetado de corpus, se han detectado un conjunto de dificultades tanto en la extracción y anotación del corpus como en el primer análisis.

En la primera fase, la *extracción* no ha acabado de ser del todo satisfactoria. Algunos de los problemas que se han detectado son los siguientes:

- el filtro de idioma no ha funcionado en algunos casos, hecho que ha dificultado el etiquetado manual.
- encontramos ejemplos repetidos y algunos que contienen publicidad, y algunas respuestas a hilos o publicaciones anteriores cuya referencia no se muestra, por lo que es necesario acudir a la publicación real para obtener las referencias necesarias para comprender el tuit; el contexto conversacional se pierde.
- el hecho de usar como primer filtro los *trending topic* de Twitter ha chocado con el uso más habitual que hacen los usuarios.
- el periodo de tiempo seleccionado para recopilar el corpus también ha provocado una mayor presencia de determinadas temáticas, tales como la política, a causa de las elecciones, o la supresión de restricciones por covid. Por tanto, no se ha conseguido equilibrar la presencia de diversas temáticas, siendo la política la más presente.

En relación con el *etiquetado*:

- Se ha detectado confusión en el término *violencia*, pues se identifican grados extremos, sin considerar los menores, como posibles casos de violencia verbal. Ello se debe a la confusión entre violencia verbal y aceptabilidad social, es decir, el hecho de que ciertos grados de violencia verbal estén permitidos y normalizados hace que no se detecten, en primera instancia, como violencia verbal.
- El hecho de no etiquetar los grados de violencia dificulta el consenso entre anotadores, pues el hecho de etiquetar cada tuit como violento o no provoca que, en los casos de bajos grados de violencia, se tienda a etiquetar como no violento.
- Se ha observado cierta sensibilidad en relación con determinadas temáticas según la persona que etiqueta (mensajes machistas etiquetados por mujeres). Este desvío se puede solventar con el comentario de estos mensajes y el acuerdo posterior, evitando la subjetividad, aunque también ofrece información útil sobre la percepción de la violencia verbal y las implicaciones que se derivan.
- Los casos que no son explícitos causan problemas a la hora de justificar lingüísticamente los elementos que contienen el contenido violento (desde estructuras concretas hasta comparaciones o enunciados de eco).

En lo referente al *análisis computacional de frecuencias*:

- Se ha observado que los diferentes usos de los *hashtags* (como etiquetas o como palabras propiamente) dificulta el análisis de ítems más frecuentes. En los análisis computacionales se suele *normalizar* el texto, eliminando los elementos no lingüísticos propiamente (emoticonos, signos de puntuación, *hashtags*). Sin embargo, el hecho de eliminar estos elementos hace que se pierda información relevante en la que pueda aparecer gran parte de la violencia del mensaje, así como la fijación de la temática o del referente.

    - En el caso de los emoticonos, pueden incrementar el nivel de violencia o incluso marcar la polaridad del mensaje (si es irónico o sarcástico, por ejemplo), así como incluir en sí un ataque a la imagen (en el caso de emoticonos escatológicos o de gestos considerados groseros).
    - En el caso de los signos de puntuación, es típica la construcción de eufemismos con varios signos o bien la intensificación mediante la repetición de signos exclamativos o interrogativos.
    - En el caso de los *hashtags*, estos pueden tener varias funciones, tales como marcar la temática, el referente o la situación a la que hace referencia, o bien incluir un insulto o construir otro mensaje dentro del mismo tuit.
- La presencia de estos *hashtags* en tuits de la misma temática altera el análisis de frecuencias, pues es común que los mensajes publicados en un mismo periodo de tiempo sobre la misma temática compartan los mismos *hashtags*.
- Además, a causa de usar como filtro los *trending topic* para extraer mensajes de diferentes temáticas, hemos detectado que muchos n-gramas se han alterado también; esto es, varios mensajes del mismo tema contienen menciones o *hashtags* parecidos, por lo que esta frecuencia de elementos puede ser relevante en relación con la información situacional.

5. DISCUSIÓN DE LOS RESULTADOS

Con el objetivo de solventar los problemas encontrados y hacer propuestas concretas, se han puesto en relación las dificultades encontradas en nuestra simulación con algunos estudios recientes en este ámbito.

5.1. El etiquetado de corpus: tipos, análisis y rasgos

Según Basile (2022), tradicionalmente, la anotación de corpus tiene un conjunto de instancias (oraciones, documentos, palabras u otras unidades lingüísticamente significativas), un fenómeno de estudio descrito en detalle, un esquema de anotación con los valores y reglas aplicables, y un grupo de anotadores. Esta fase de etiquetado, en el ámbito computacional, se puede realizar en un primer momento manual o automáticamente. La anotación manual es la que actualmente tiene más presencia, pues el etiquetado automático requería una base de datos explícita de la que partir, por lo que contemplaba los términos como violentos intrínsecamente y, por tanto, se etiquetaban únicamente los mensajes explícitos. Asimismo, el etiquetado manual se puede realizar con una guía de anotación (Díaz-Torres y otros, 2020) o sin esta. A pesar de que van ganando presencia las investigaciones que desarrollan guías de anotación, todavía son habituales las que no las utilizan.

En lo referente a los anotadores, estos pueden ser expertos o *crowd*. Los expertos suelen ser investigadores del fenómeno de estudio, aunque también se incluyen víctimas de *hate speech* o activistas por los derechos sociales. La anotación *crowdsourcing*, sin embargo, se realiza a través de plataformas como Amazon Mechanical Turk, mediante las que se realiza una selección de los anotadores (Basile, 2022). Desconociendo la identidad y características de los etiquetadores, un número pequeño de ejemplos es etiquetado mediante una guía de anotación que sirve de filtro según el acuerdo que los resultados de esta anotación muestren. Así, las personas

con mayor acuerdo son las finalmente seleccionadas para la tarea de etiquetado del corpus.

Sin embargo, las guías de anotación suelen poseer algunos problemas definitorios, hecho causado por la ausencia de consenso en el concepto (Waseem y otros, 2017): mezclan aspectos lingüísticos y sociales (nivel de explícito con nivel de ataque) y no consideran la información situacional y contextual. De este modo, las clasificaciones de categorías de violencia verbal tampoco son claras: mezclan categorías o recursos lingüísticos con actividades de imagen (insultos y sarcasmo) (Assimakopoulos y otros, 2020), así como términos sin definición clara y distintiva (*hate speech*, ofensivo) (Ombui y otros, 2021). Estos problemas terminológicos provocan resultados dispares y no generalizables.

En relación con el desacuerdo, algunos autores apuestan por analizarlos en lugar de eliminarlos (Sang y Stanton, 2021). A pesar de que se suelen *harmonizar* los resultados, esto es, comentar aquellos ejemplos con etiquetados distintos con el objetivo de consensuar un etiquetado definitivo, el desacuerdo puede ofrecer información útil en el estudio de fenómenos como el *hate speech*.

Sang y Stanton (2021), por ejemplo, afirman que los errores o desacuerdos en el etiquetado se dan porque se asume que hay una única anotación universal, que considera categorías excluyentes. Es por ello por lo que pueden darse anotaciones dispares y el hecho de que las percepciones de los hablantes sean distintas es porque influyen varios factores como el género y la etnia de los propios anotadores. Así, diferentes anotadores pueden dar luz a distintos esquemas mentales que intervienen en la comprensión y que dan lugar a distintas anotaciones (Sang y Stanton, 2021).

Basile y otros (2021: 16), por su parte, explican que la eliminación del desacuerdo puede ofrecer unos resultados de evaluación mejores, pero esconden la verdadera naturaleza de la actividad que se pretende solucionar. Cabe distinguir, pues, las anotaciones 'objetivas' aquellas cuyo objeto de anotación puede ser un fenómeno lingüístico, de las 'subjetivas', relacionadas con fenómenos sociales como el *hate speech*. A pesar de que el desacuerdo se puede dar en ambas, es en este último tipo de anotación que puede ser verdaderamente valiosa. De este modo, al huir del *gold standard* —conjuntos de datos anotados manualmente por expertos y que se toman como referencia (Wissler y otros 2014) o anotación de cada ítem tomada como correcta (Novak y otros, 2022)—, se consideran las anotaciones individuales, sin ser ninguna errónea, que permiten crear modelos inclusivos con respecto a los conocimientos subjetivos que intervienen en la comprensión de los mensajes (Basile y otros 2021: 18).

De este modo, Novak y otros (2022) exponen que, en contraste con el *gold standard*, ha surgido el llamado *diamond standard*, que huye del etiquetado único y ofrece un amplio abanico de posibilidades de etiquetado. Así, cuantos más prismas se consideren, esto es, cuantas más opiniones diferentes entre anotadores, mayor calidad de los datos (Novak y otros, 2022: 691). De este modo, los resultados pueden ofrecer una visión más realista y, por tanto, una aplicación más cercana al razonamiento y comprensión humanas.

Referente a la fase de análisis computacional propiamente, en el análisis del *hate speech* se suelen extraer frecuencias de elementos generalmente léxicos y sintácticos, aunque seguimos encontrando dificultades con el uso de extracción de frecuencias en diferentes niveles lingüísticos (Fortuna y Nunes, 2018). No se incluyen, pues, informaciones situacionales y contextuales, elementos fundamentales en la evaluación de la violencia verbal (Fernández, 2016).

Los análisis computacionales realizados mediante los algoritmos tradicionales, los rasgos que se consideran en la clasificación automática son accesibles, por lo que los investigadores pueden observar qué características se han considerado en esta clasificación. Sin embargo, en los algoritmos basados en redes neuronales dicha información no es accesible, de modo que no es posible conocer los rasgos distintivos de las diferentes clases consideradas.

Además, no se suele realizar un análisis lingüístico posterior que pueda determinar qué rasgos o atributos son más representativos del fenómeno en cuestión. Esta revisión manual es necesaria, pues la contribución de un experto en el análisis computacional permite afinar más el análisis realizado y, por tanto, poder obtener unos resultados fundamentados y explicables.

5.2. Algunas propuestas de solución

Partiendo de los problemas en el etiquetado y el análisis, planteamos un conjunto de soluciones que pueden solventarlos.

En primer lugar, el etiquetado de mensajes violentos y de los grados de violencia requiere de una guía de etiquetado clara, comentada y consensuada, así como el comentario posterior para garantizar un mayor acuerdo y coherencia entre ejemplos. Consideramos la violencia verbal como un fenómeno difuso, esto es, que contempla una serie de grados entre los que puede haber mayor o menor aceptación social (desde el uso de groserías hasta la categorización de la persona o el rechazo de sus acciones). Estos grados posiblemente se correspondan con diferentes acciones pragmáticas o actividades de imagen que se llevan a cabo mediante diferentes estrategias, comunicadas de forma tanto explícita como implícita.

Es en esta distinción de mensajes explícitos e implícitos en los que radica la dificultad del análisis computacional. Por ello, es necesario llevar a cabo un etiquetado posterior realizado por expertos en el que se anote cada ejemplo como explícito o implícito, considerando diferentes implicaturas atendiendo a la información más presente lingüísticamente (desencadenantes de implicaturas o *triggers*). También es imprescindible que se realice un análisis manual posterior en el que se distingan los distintos tipos de implicaturas y en los que se marquen los elementos que las desencadenan o la información, tanto situacional como lingüística, que en conjunto permite comprender cada mensaje.

Asimismo, cabe considerar también el desacuerdo. Atendiendo a un fenómeno sociopragmático como la violencia verbal, es necesario tener en cuenta la comprensión que los hablantes tienen del fenómeno. Este desacuerdo nos permitirá, junto con las características de los anotadores, establecer relaciones entre rasgos personales y percepción de violencia que pueden ser interesantes. Además, la comunicación en las redes sociales está en cambio constante, por lo que la percepción de la violencia verbal en Twitter también puede variar. La mayor presencia de mensajes violentos puede provocar que se normalice la lectura de estos mensajes y, por ende, puede resultar en una menor percepción consciente o una mayor tolerancia.

En lo que al análisis computacional en sí se refiere, es necesario tener en cuenta los recursos empleados en la red. El análisis del lenguaje natural engloba el estudio del lenguaje en cualquiera de sus vertientes, por lo que ha de tomar en consideración los recursos empleados en las redes sociales que intervienen en la comprensión de mensajes. Elementos como los *hashtags* y los emoticonos ofrecen información relevante para la comprensión de los mensajes con violencia verbal, por lo que es necesario considerarlos en el análisis.

Además, cabe distinguir los rasgos o atributos más frecuentes de los más representativos del fenómeno. Es necesario diferenciar los resultados que muestren características propias del género en la red social y los resultados propios de los mensajes violentos. Esto tiene en cuenta la diferencia del género y registro en Twitter, así como las limitaciones de caracteres que la red impone. Esta limitación conlleva una mayor 'condensación' de la información disponible en la emisión y recepción de mensajes, por lo que los recursos que se emplean pueden ser variados. La variedad de recursos utilizados en la red no se corresponde, pues, con una mayor frecuencia de determinadas estructuras lingüísticas, y es la combinación de rasgos o atributos tanto lingüísticos como situacionales la que puede solventar, en parte, esta limitación.

6. GUÍA DE ETIQUETADO: UNA PROPUESTA

Con el objetivo de realizar el etiquetado de nuestro corpus completo, que comprende 3.000 tuits, y analizarlo computacionalmente, se ha elaborado, por un lado, una guía de anotación y, por otro lado, una lista de rasgos que pueden ser prototípicos de la violencia verbal en Twitter.

6.1. Guía de anotación

La guía de anotación incluye nuestra definición de violencia verbal y la explicación de las premisas socioculturales propias del contexto español. Junto con estas definiciones, se plantean dos preguntas con el objetivo de guiar el etiquetado.

El concepto de violencia verbal del que partimos es genérico, esto es, incluye diferentes tipos de violencia verbal y diferentes grados, tales como la descortesía o el *hate speech*. Esta definición bebe de los elementos comunes a las definiciones de violencia de distintos ámbitos (Henry y Milovanovic, 1996; Levi y Maguire, 2002; Riches, 1986) y de las definiciones de descortesía, término empleado tradicionalmente en el ámbito de la lingüística para hacer referencia a la violencia verbal. Con esta definición evitamos emplear términos de categorías como *insulto* o *amenaza*, que puedan generar disparidad de comprensiones.

Asimismo, se exponen los conceptos sociopragmáticos básicos para fundamentar el concepto de violencia verbal y garantizar la coherencia entre etiquetado y concepto del que se parte. Se definen términos como 'premisa sociocultural', 'imagen social', 'actividad de imagen' y 'efecto social', partiendo de Bravo (1999, 2004).

Sin embargo, para obtener una mirada amplia y variada de las comprensiones de violencia verbal, se pide a los anotadores que etiqueten cada tuit en una escala Likert del 1 al 7, siendo 1 poco y 7 mucho, según el grado de violencia que perciban. De este modo, podremos observar, en tres niveles mayores (poco, bastante, mucho), la aceptación social de ciertos mensajes y el acuerdo o desacuerdo correspondiente a cada uno. Los rasgos de los anotadores nos ofrecerán información que influye en la comprensión de dichos mensajes.

La información que se muestra en la guía es la siguiente:

- Violencia verbal: acto realizado mediante el lenguaje que daña la imagen (autoconcepto) del receptor o de la persona a la que hace referencia. Se basa en las normas sociales y se percibe como intencional.

    - ¿Considera que el tweet puede dañar la imagen de una o varias personas? En el caso de que fuera dirigido a usted, ¿se sentiría atacado/a? No = 0, Sí = 1. Si responde "No", salte al siguiente tweet.
- Premisa sociocultural: socialmente se comparte que se ha de respetar la imagen y contribuir a su reafirmación.
- Imagen social: afirmación de la originalidad y de las buenas cualidades; reciprocidad y generosidad, respeto por la posición social.
- Actividad de imagen: acción relacionada con afectaciones positivas, negativas o neutras sobre la imagen de los interactantes.
    - Efecto social: la comunicación tiene un efecto básico en la identidad psicosocial de la persona, en su imagen social. En una escala de 1 (poco) al 7 (mucho), ¿en qué grado es violento dicho tweet?

6.2. Lista de rasgos lingüísticos

Nuestro análisis computacional se va a llevar a cabo mediante una lista de atributos en diferentes niveles lingüísticos, desde elementos gráficos hasta elementos situacionales. Se han tenido en cuenta las características de la red, por lo que se consideran aquellos rasgos que puedan aparecer o que ofrece Twitter. Este listado contempla:

- Elementos gráficos
    - Caracteres especiales (#%<¨)
    - Signos de puntuación
    - Letras repetidas
    - Emoticonos (sobre todo, los referentes a emociones y otros propios de mensajes violentos como 🔪 💩 🖕)
- Onomatopeyas
- Interjecciones e interpelaciones
- Elementos morfológicos
    - Diminutivos
    - Aumentativos
    - Peyorativos
    - Superlativos)
- Léxicos
    - Artículos determinados, indeterminados y demostrativos
    - Pronombres personales
    - Verbos atributivos
    - Verbos en su distinta flexión de tiempo, modo y persona
    - Adverbios de cantidad, negación y duda
    - Adjetivos de cantidad y peyorativos
    - Palabras relacionadas con acciones escatológicas
    - Insultos
- Construcciones sintácticas
    - Verbos (+ Adverbio) + Adjetivo
    - Verbo copulativo (+ Adverbio) + Adjetivo
    - Verbo + Determinante + Sustantivo (+ Adjetivo)
    - Construcciones groseras (adjetivo y *de mierda*…)
    - Oraciones comparativas y desiderativas;

- Elementos situacionales
    - Nombre de la cuenta
    - Nombre de usuario
    - Número de seguidores
    - Número de *retweets*, *replies* y *me gusta*
    - Menciones y *hashtags*).

De este modo, el análisis computacional considerará tanto la presencia de dichos atributos como la frecuencia y combinación de rasgos en cada mensaje. Para valorar la representatividad de dichos rasgos, se compararán dos tipos de algoritmos: el aprendizaje no supervisado mediante el corpus sin etiquetar y el aprendizaje supervisado mediante el corpus etiquetado. Esta comparativa permitirá valorar si la presencia y frecuencia de rasgos es propia del fenómeno o de los mensajes de Twitter en general.

7. CONCLUSIONES

Tanto la importancia del análisis lingüístico como del computacional es irrefutable en el estudio de la violencia verbal en la red. Ambas disciplinas ofrecen posibilidades y fundamentos que son necesarios para la teorización del fenómeno como para su detección automática. A pesar de que actualmente los estudios en ambas ramas están en incremento, todavía encontramos problemas similares que no se han conseguido resolver y que influyen en los resultados que se obtienen.

En los dos ámbitos que se ocupan de la violencia verbal, la lingüística y la computación, observamos que los problemas surgen principalmente por la disparidad de definiciones de términos como *hate speech*. Las categorías o subtipos de descortesía o *hate speech* también son varias y no se establecen con unos criterios claros. Es destacable también la diferencia sociocultural en la concepción de la violencia verbal, hecho que hace necesaria la consideración del contexto en la evaluación del fenómeno. Además, se concibe la violencia verbal como un fenómeno discreto, por lo que los etiquetados de corpus que se realizan se limitan a una respuesta 'correcta', sin considerar el desacuerdo entre anotadores como fuente de información de la comprensión de dichos mensajes. Asimismo, los análisis computacionales se centran en la extracción de frecuencias, a pesar de que puedan obtener rasgos característicos de la red y no de los mensajes violentos propiamente.

Con el objetivo de abordar estos problemas y proponer algunos cambios, hemos realizado una simulación de etiquetado de un corpus de Twitter en términos de violencia verbal. Defendemos que la violencia verbal es un fenómeno difuso, con grados, y que engloba varios subtipos de violencia. Tras la simulación realizada, creemos necesario elaborar una guía de etiquetado que permita obtener resultados coherentes con la premisa o concepto del que se parte. También se debería tomar en consideración el desacuerdo entre anotadores como fuente de información relevante para el estudio de la violencia verbal, pues es un fenómeno que, junto con las redes sociales, está en constante cambio. Asimismo, se debe de realizar un análisis de dichos mensajes en términos de explícito-implícito, pues la violencia verbal no siempre está presente lingüísticamente, análisis que debe llevar a cabo un experto por los matices que pueda tener esta categorización. Puesto que esta distinción es difícil de capturar en los análisis de frecuencias con ítems lingüísticos, es necesario apostar por el análisis de recursos y frecuencias también en otros niveles lingüísticos

(sintáctico, semánticos y pragmáticos). Por tanto, en el análisis computacional es necesario considerar los rasgos del fenómeno de estudio para obtener resultados generalizables y característicos, rasgos que se obtendrán gracias al análisis lingüístico.

En síntesis, el análisis de la violencia verbal necesita de la conexión entre la lingüística y la computación para poder realizar análisis eficaces, fundamentados, y obtener resultados explicables y generalizables. La figura de lingüista es imprescindible para solventar los problemas en el procesamiento del lenguaje natural que el análisis computacional no ha podido solucionar.

REFERENCIAS

Assimakopoulos, S., Vella Muskat, R., van der Plas, L. y Gatt, A. 2020. Annotating for hate speech: The MaNeCo Corpus and some input from Critical Discourse Analysis. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 5088–5097.

Basile, V. 2022. The perspectivist data manifesto. <https://pdai.info> (acceso: 17/08/2022)

Basile, V. Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M. y Uma, A. 2021. We need to consider disagreement in evaluation. *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, 15–21.

Bou-Franch, P. 2021. 'Maleducados/Ill-mannered' during the #A28 political campaign on Twitter: A metapragmatic study of impoliteness labels and comments in Spanish. *Journal of Language Aggression and Conflict* 9(2): 271–296.

Bravo, D. 2004. Tensión entre universalidad y relatividad en las teorías de cortesía. En D. Bravo y A. Briz eds. *Pragmática sociocultural: Estudios del discurso de cortesía en español*. Barcelona: Ariel, 15–33.

Bravo, D. 1999. ¿Imagen 'positiva' vs. imagen 'negativa'? Pragmática sociocultural y componentes de face. *Oralia: Análisis del discurso oral* 2: 155–184.

Brown, P. y Levinson, S. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Culpeper, J. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics* 25: 349–367.

Culpeper, J. 2017. Impoliteness metalinguistic labels and concepts in English. En R. Giora y M. Haugh eds. *Doing pragmatics interculturally. Cognitive, philosophical and sociopragmatic perspectives*. Berlín: Mouton de Gruyter, 135–147.

Díaz-Torres, M. J., Morán-Méndez, P. A., Villaseñor-Pineda, L., Montes-y-Gómez, M., Aguilera, J. y Meneses-Lerín, L. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 132–136.

Fernández, F. 2016. Bases teóricas para un estudio transcultural y variacionista de la (des)cortesía. *ELUA* 30: 79–100.

Fortuna, P. 2017. Automatic detection of hate speech in text: An overview of the topic and dataset annotation with hierarchical classes. Trabajo de Fin de Máster, Universidad de Oporto.

Fortuna, P. y Nunes, S. 2018. A survey on automatic detection of hate speech. *ACM Computing Surveys* 51(4): 1–30.

Goffman, E. 1955. On face work. An analysis of ritual elements in social interaction. *An Analysis of Ritual Elements in Social Interaction* 18(3): 213–231.

Henry, S. y Milovanovic, D. 1996. *Constitutive criminology. Beyond postmodernism*. Londres: Sage.

Levi, M. y Maguire, M. 2002. Violent crime. En M. Maguire, R. Morgan y R. Reiner eds. *The Oxford handbook of criminology*. Oxford: Oxford University Press, 687–732.

Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozetič, I. y Zollo, F. 2022. Handling disagreement in hate speech modelling. *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 681–695.

Ombui, E., Muchemi, L. y Wagacha, P. 2021. Building and annotating a codeswitched hate speech corpora. *International Journal of Information Technology and Computer Science* 3: 33–52.

Plaza-del-Arco, F. M., Molina-González, M. D.; Ureña-López, L. A. y Martín-Valdivia, M. T. 2020. Detecting misogyny and xenophobia in Spanish tweets using language Ttechnologies. *ACM Transactions on Internet Technology* 20(2): 1–19.

Poletto, F. Basile, V., Sanguinetti, M., Bosco, C. y Patti, V. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Lang Resources & Evaluation* 55: 477–523.

Riches, D. 1986. *The anthropology of violence*. Oxford: Blackwell.

Sang, Y. y Stanton, J. 2021. The origin and value of disagreement among data labelers: A case study of the individual difference in hate speech annotation. *iConference 2022: Information for a Better World: Shaping the Global Future*, 425–444.

Scheper-Hughes, N. y Bourgois, P. 2004. Introduction: Making sense of violence. En N. Scheper-Hughes y P. Bourgois eds. *Violence in war and peace*. Oxford: Blackwell, 1–31.

Terkourafi, M. 2008. Toward an unified theory of politeness, impoliteness, and rudeness. En D. Bousfield y M. A. Locher eds. *Impoliteness in language: Studies on its interplay with power in theory and practice*. Berlín: Mouton de Gruyter, 45–76.

Waseem, Z., Davidson, T., Warmsley, D. y Weber, I. 2017. Understanding abuse: A typology of abusive language detection subtasks. *Proceedings of the First Workshop on Abusive Language Online*, 78–84.

Wiegand, M., Ruppenhofer, J. y Eder, E. 2021. Implicitly abusive language – What does it actually look like and why are we not getting there? *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 576–587.

Wissler, L., Almashraee, M., Monett, D. y Paschke, A. 2014. The gold standard in corpus annotation. *5th IEE Germany Sutdent Conference*, 1–4.