¿ES POSIBLE LA DETECCIÓN AUTOMÁTICA DE LA VIOLENCIA LINGÜÍSTICA EN LAS REDES SOCIALES?

Susana Campillo Muñoz
Universitat Rovira i Virgili, Tarragona

RESUMEN

La violencia lingüística es un fenómeno presente en la actualidad. En las redes sociales ve modificada la forma en que se materializa y esto ha impulsado la creación de métodos computacionales para su análisis automatizado. Esta investigación se centra en el análisis de enunciados violentos en el contexto virtual. Observamos que en los desarrollos computacionales actuales no se limitan las categorías de violencia, se emplean diferentes clases según los autores y se consideran categorías discretas. Además, son escasos los trabajos que consideran la información extralingüística y los recursos propios del mundo virtual. Por ello, defendemos la inclusión de la información contextual, tanto sociocultural como de la red y del usuario, para desarrollar una categorización en términos difusos y en diferentes niveles lingüísticos. Creemos necesario el trabajo conjunto de informáticos y lingüistas.

Palabras clave: violencia lingüística, detección automática, redes sociales

ABSTRACT

Verbal violence is a phenomenon with a strong impact on contemporary society. In social media, however, the way it emerges differs, propelling thus the formulation of computational analysis methods to detect it and prevent it. This investigation focuses on violent statements analysis in a virtual context. The categories used in current computational detection systems are not limited, show differences across authors and are treated as discrete entities. In addition, there are few studies that include extralinguistic data and virtual communication features. Therefore, in this paper I advocate the inclusion of sociocultural, network and user contextual data, in order to develop a classification system which uses fuzzy categories and various linguistic levels. This calls for the collaboration of linguists and computer engineers.

Keywords: linguistic violence automatic detection, social media

1. INTRODUCCIÓN

La violencia verbal y, en concreto, en las redes sociales es un fenómeno presente en nuestra sociedad. Las redes sociales han conformado y forman un entorno virtual en el que las relaciones sociales se ven modificadas.

Con relación a la violencia verbal, cada día son más numerosas las noticias de denuncias a cantantes, humoristas y otros profesionales por enunciados emitidos en el contexto virtual. Estos hechos derivan en diversidad de opiniones de los hablantes sobre estos sucesos y sobre las condenas que se imponen. Asimismo, la variabilidad de consecuencias legales que conllevan los enunciados violentos en general, y aún más en internet, refleja la importancia del componente contextual y sociocultural en sus valoraciones.

El interés y la actualidad del fenómeno han impulsado a los investigadores a desarrollar herramientas para el análisis de estos enunciados, catalogándolos bajo etiquetas diferentes: *hate speech*, *ciberacoso, lenguaje racista, sexista u ofensivo*, entre otros. No obstante, nos seguimos encontrando con casos difíciles de analizar y de clasificar, como los enunciados implícitos.

A partir de lo dicho, nos preguntamos si es posible la detección automática de estos enunciados atendiendo a la variabilidad inherente del fenómeno.

Nuestra investigación parte de la pragmática y se centra en los enunciados violentos en un contexto determinado, en este caso, en el entorno virtual.

El objetivo principal es proponer una formalización que dé cuenta de la información relevante para el análisis de los mensajes violentos. Para ello, nos fijamos los siguientes subobjetivos:

- Definir la violencia verbal y establecer una clasificación en términos difusos de las distintas categorías de violencia verbal.
- Delimitar los patrones recurrentes en los enunciados violentos en las redes sociales.
- Concretar la información lingüística y sociocultural que influye en la comunicación de la violencia verbal.
- Establecer el modo más adecuado de introducir la información lingüística y sociocultural en su análisis.

Así pues, por un lado, estudiaremos tanto la bibliografía existente sobre violencia verbal como el conocimiento metapragmático de los hablantes; por otro lado, analizaremos los enunciados violentos de un corpus extraído de las

redes sociales, y las investigaciones y desarrollos computacionales referentes a dicho fenómeno.

2. MARCO TEÓRICO

El fundamento teórico de nuestro estudio parte de cuatro teorías que se toman como base y de diferentes estudios en relación a la violencia verbal. Así, nuestra investigación parte, en primer lugar, de la teoría de los actos de habla (Austin, 1955), puesto que afirmamos que al hablar se realiza un tipo particular de acción; de la teoría de la cortesía (Brown y Levinson, 1987) y de su contraparte, la teoría de la descortesía, (Culpeper, 1996, 2010; Culpeper, Haugh, Sinkeviciute), ya que la violencia verbal combina la parte lingüística y la social; y de la teoría de la relevancia (Sperber y Wilson, 1994), que nos permite calcular el grado de implícito y explícito atendiendo a la diferente información contextual.

Partimos de las siguientes premisas teóricas en relación con la violencia verbal. El fundamento de la violencia es cultural y, por tanto, el grado de variabilidad determina la diversidad de marcos contextuales desde los que se evalúa. Además, la violencia verbal se puede realizar de manera explícita o implícita [el 34% de los tweets irónicos son ataques, ridiculizan o dañan a alguien (Van Hee et al., 2018)]. Por lo que se refiere a la evaluación y valoración de la violencia verbal, esta es contextual, es decir, la información situacional y del contexto comunicativo es imprescindible para analizarla.

Por otro lado, las redes sociales suponen un cambio de paradigma en los estudios de comunicación (Herring, 2004). Tanto la información disponible para el usuario, como el contexto de comprensión y los recursos comunicativos son distintos a los de la comunicación cara a cara; han surgido, pues, nuevos géneros textuales con características formales diferentes que hay que considerar para un análisis automático eficaz.

Resumiendo, las redes sociales son un espacio en el que la violencia lingüística tiene lugar con características y significado social modificados respecto a su realización en el mundo físico (Miró, 2016). La lingüística computacional ofrece herramientas y perspectivas desde las que comprobar las diversas teorías sobre la comunicación mediada por ordenador. El procesamiento del lenguaje natural para las redes sociales requiere adaptar los métodos tradicionales o crear nuevos (Farzindar, 2018).

Parece claro que, para la detección automática de la violencia verbal en las redes sociales, es necesario un cambio de perspectiva en referencia al paradigma tradicional del estudio tanto de la comunicación como de la violencia verbal.

3. PROBLEMAS EN EL ANÁLISIS DE LA VIOLENCIA VERBAL

Tras realizar un estudio de la bibliografía existente y de las investigaciones actuales sobre el tema, hemos detectado un conjunto de problemas, tanto teóricos como de la aplicación práctica en la detección de la violencia verbal. Clasificamos estos problemas en dos bloques: problemas teóricos y problemas computacionales, aunque tienen una clara relación y se retroalimentan.

3.1 Problemas teóricos

Entre las investigaciones actuales encontramos varios trabajos de detección de mensajes racistas, *hate speech*, ciberacoso y lenguaje ofensivo. Así, observamos distintas materializaciones del fenómeno, es decir, varios subtipos que se engloban bajo un término genérico como puede ser *descortesía* o *agresión verbal* (Culpeper, 2017), y encontramos también una escasa delimitación entre estas clases. No se distinguen las etiquetas empleadas en las categorías y, de hecho, se emplean clasificaciones distintas según el ámbito de estudio del que se parta (Schmidt y Wiegand, 2017).

Asimismo, el contexto comunicativo en el mundo virtual posee unas características diferentes que afectan a la calidad y cantidad de información contextual (Yus, 2018). Surgen nuevos tipos de textos o géneros textuales de las diferentes herramientas de comunicación en línea, y estos textos poseen unos rasgos recurrentes y unas características lingüísticas propias. Por tanto, parece lógico que se tenga en cuenta, en el análisis de la violencia verbal, la información contextual compartida, que dependerá de la red social en cuestión y de la información que incorpore, y el género discursivo que se investiga. Esta información es imprescindible a la hora de determinar qué rasgos considerar en su análisis automático (Brenes Peña y Fuentes Rodríguez, 2017) y también cómo procesarlos para una posterior detección automática. Estudios recientes empiezan a incorporar esta información (Schmidt y Wiegand, 2017); sin embargo, no se aprecia en general una

inclusión de las características del contexto comunicativo en la mayoría de las investigaciones.

De este modo, los problemas teóricos expuestos se convierten en metodológicos. El hecho de la falta de caracterización clara y uniforme de la violencia verbal conlleva distintas metodologías de análisis (Ajdari y Ismaili, 2017) según se conciba como un fenómeno gradual o con categorías cerradas, por ejemplo. Las clasificaciones, además de ser varias, parten de distintos parámetros (la intención, el grado de lesión…), hecho que dificulta un análisis minucioso con las herramientas computacionales disponibles.

Todo esto aplicado al análisis automático conlleva una concepción simplista de la violencia verbal, centrada en el análisis del mensaje atendiendo básicamente a la parte textual (Ajdari y Ismaili, 2017) y, por tanto, resulta ser un análisis parcial de dicho fenómeno.

3.2 Problemas computacionales

En los desarrollos computacionales las categorías de violencia verbal que se emplean también varían según los autores (Waseem, 2016; Razavi et al*.,* 2010), no se definen claramente y se consideran categorías discretas.

Se analiza, pues, la presencia o ausencia de palabras consideradas obscenas, racistas u ofensivas. Además, se emplean métodos estadísticos y la mayoría de los estudios consideran enfoques generales, sin atender a rasgos particulares para la violencia verbal como sucede, por ejemplo, en el discurso de odio con la apelación de un «vosotros» frente a un «nosotros» (Fortuna, 2017; Biere, 2018).

A esto hay que añadir las diferencias entre la información que se considera y los niveles de análisis. Son pocos los trabajos que incluyen la información extralingüística (información del usuario, de la red social...) y los recursos de estas plataformas (puntuación, emoticonos…), y abundan los análisis léxico-sintácticos (Burnap y Williams, 2016): la minería del texto basada en frecuencias de tokens, bolsas de palabras (Nahar, Li y Pang, 2013) o analizadores sintácticos.

No obstante, si se pretende identificar estos mensajes es esencial considerar las condiciones contextuales (Whitten, 2018). Algunos estudios recientes empiezan a incluir más niveles de análisis e información extralingüística,

tales como los estados sociales, las emociones (Douce Munezero, 2017) o el análisis de opinión y la polaridad; otros introducen también el conocimiento del mundo y la metainformación del usuario (Schmidt y Wiegand, 2017). Cabe destacar la importancia de los límites contextuales que subrayan algunos autores para identificar el lenguaje grosero como recurso expresivo (Biere, 2018), sin considerarlo violencia verbal.

Con relación al procesamiento de los mensajes, en la fase de normalización del texto se excluyen las marcas de puntuación propias de recursos estilísticos, emoticonos, mayúsculas…, con lo que consideramos que se pierde información relevante. Las clasificaciones son binarias (violento o no a partir de la asignación de violencia inherente en el término) y se aplican clasificaciones de otros contextos socioculturales o traducciones literales de estudios en otros géneros y entornos culturales.

Por último, observamos variedad de algoritmos y de distinto tipo de aprendizaje: supervisado (Douce Munezero, 2017) y no supervisado (Gordeev, 2016); métodos clásicos y aprendizaje profundo (Badjatiya et al., 2017).

Consideramos, así, que los lingüistas e investigadores deberían ir a la par con la evolución de las tecnologías digitales y de las relaciones sociales que se establecen en ellas (Androutsopoulos y Beißwenger, 2008).

4. PROPUESTA PARA EL ANÁLISIS DE LA VIOLENCIA VERBAL

Tras resumir los distintos problemas teóricos y prácticos para el análisis de la violencia verbal en las redes sociales, proponemos lo siguiente.

En primer lugar, es necesaria una descripción y caracterización clara del fenómeno para poder desarrollar una formalización que dé cuenta de la información relevante para su detección automática. Respecto a las clasificaciones, consideramos que es necesario atender a la variabilidad contextual, tanto en lo referente a la parte sociocultural como a la de la situación comunicativa. Para ello, hay que tener en cuenta tanto la información bibliográfica de las teorías tradicionales como la de las investigaciones actuales, así como el conocimiento de los hablantes con relación al fenómeno. Se ha de considerar el marco contextual y sociocultural en el caso del español, como ya hacen algunos estudios sobre la comprensión de la ironía, atribuyendo, por ejemplo, marcos a palabras, a cadenas de palabras y a situaciones (Yus, 2013).

Cabe considerar, además, la posible diferente percepción del mismo

fenómeno en el contexto virtual y el contexto físico: atendiendo a la mayor exposición de la imagen positiva en el mundo virtual, ¿un mismo acto realizado en internet y cara a cara es violento en la misma medida en ambos contextos? Es necesario también tener en cuenta estos factores y clasificar los mensajes según el grado de violencia en su emisión a través de las redes sociales.

Como hemos expuesto, las redes sociales conforman un entorno en el que la información disponible en el proceso de comunicación se ofrece de manera distinta. Es pertinente, pues, tener en cuenta qué información se tiene al alcance al emitir y recibir estos enunciados violentos y cómo se puede formalizar. Se ha de atender al contexto compartido en el mundo virtual estableciendo, por ejemplo, condiciones de *tipicalidad* que caractericen el prototipo del fenómeno [algunos autores, como Utsumi (2007), emplean esta metodología en el estudio de la ironía].

Finalmente, la violencia verbal se puede comunicar de manera implícita o explícita; por tanto, se han de incluir diferentes niveles lingüísticos tanto en la extracción de patrones como en el análisis de enunciados. Creemos necesario incluir diferentes rasgos en diferentes niveles [puntuación, emoticonos, mayúsculas… (Robinson et al., 2018)] y aquellos fenómenos de *citacionalidad* (Butler, 1997) que evocan situaciones violentas.

Por tanto, proponemos:

- Caracterizar de manera clara, no ambigua, la violencia verbal en el contexto español. Se analizarán las definiciones tradicionales de descortesía realizadas en el ámbito lingüístico y las del ámbito legal, fijando una caracterización concreta para el contexto cultural español.
- Establecer categorías difusas en las clasificaciones de violencia verbal. Las diferentes clases que forma parte del fenómeno son graduales, por lo que se han de fijar en un continuum. Estas categorías se extraerán de los estudios realizados y del conocimiento metapragmático de los hablantes mediante encuestas.
- Realizar un análisis en diferentes niveles lingüísticos. Se extraerá un corpus de datos y se estudiarán los rasgos lingüísticos (fonéticos, léxicos, sintácticos y semánticos) propios de las categorías anteriormente fijadas, así como los recursos más comunes (metáforas, comparaciones…).
- Considerar la información disponible en la red social (textual, del usuario, de la red…), siendo conscientes de las limitaciones por motivos

de privacidad. Analizaremos los patrones situacionales que se repiten en los enunciados de violencia verbal para fijar las variables estables en estos enunciados.

- Extraer patrones lingüísticos y pragmáticos (hablantes, tema, contexto comunicativo, recursos) de la violencia verbal. Una vez investigados los rasgos lingüísticos y situacionales, formalizaremos estas características en las categorías de violencia verbal para posibilitar su análisis computacional.

Para ello, será necesario analizar un corpus de datos que se extraerá de las redes sociales (con herramientas como la API de Twitter) y analizar también el discurso de los hablantes sobre el fenómeno (mediante estudios psicolingüísticos) y los estudios realizados en el ámbito académico.

5. CONCLUSIONES

En este trabajo nos hemos centrado en la violencia verbal, en concreto en los problemas que hemos detectado en su análisis y hemos propuesto algunas modificaciones para el análisis del fenómeno en las redes sociales. Destacamos tres puntos fundamentales de nuestra investigación.

En primer lugar, la violencia verbal es cultural y es necesario incluir la información sociocultural en su análisis. En segundo lugar, las categorías de violencia verbal difieren entre culturas y suponen diferentes grados de violencia, por lo que es necesaria una formalización de estas categorías en términos difusos y enmarcadas en su contexto sociocultural. En tercer lugar, las redes sociales ofrecen recursos y herramientas diferentes a las de la comunicación cara a cara. Parece lógico modificar la perspectiva de análisis e incorporar estos recursos lingüísticos en la detección de mensajes violentos en las redes sociales.

Los métodos de detección automática deberían considerar las tres premisas anteriores y ampliar la información que incorporan en el análisis de estos mensajes: incluyendo la información sociocultural y contextual, la del género textual en cuestión y la variabilidad en la percepción del fenómeno.

Creemos necesario revisar y actualizar las bases teóricas de la violencia verbal y de la información que incluyen los desarrollos computacionales. Para poder desarrollar un análisis computacional de la violencia verbal es crucial examinar la información sociocultural relevante y los diferentes niveles lingüísticos, teniendo en cuenta el contexto que suponen las redes sociales. Es necesario, pues, el trabajo conjunto de lingüistas e informáticos.

Por todo lo dicho, el trabajo que estamos realizando se centra, por un lado, en la investigación de las variables situacionales o *condiciones idóneas* (Utsumi, 2007) constantes en la emisión de estos enunciados para delimitar el marco situacional y, por otro lado, en el estudio de los diferentes métodos computacionales existentes y aplicados en la detección del fenómeno.

Asimismo, estamos trabajando en la recopilación del corpus: considerando el método más adecuado, atendiendo a diferentes niveles lingüísticos, a la información disponible (de la red, del usuario, del conocimiento del mundo) y las restricciones legales. Recopilado el corpus, llevaremos a cabo la extracción de rasgos y patrones de estos enunciados que nos permitirá proponer una formalización de la información a considerar en la detección automática de la violencia verbal en las redes sociales.

REFERENCIAS BIBLIOGRÁFICAS

Ajdari, J., Ismaili, F. 2017. "Automatic Hate Speech Detection in Online Contents Using Latent Semantic Analysis". 2nd World Conference on Technology, *Innovation and Enterpreneurship*. 12-14 de Mayo, Estambul, Turquía.

Androutsoupoulos, J.; Beißwenger, M. 2008. "Introduction. Data and Methods", Computer-Mediated Discourse Analysis. Language@Internet. 5(9): 1-7.

Austin J. 1955. *Cómo hacer cosas con palabras: palabras y acciones.* Barcelona: Paidós, 1971.

Badjatiya, P., Gupta, S., Gupta, M. 2017. "Deep Learning for Hate speech detection in tweets", *Proceedings of the 26th International Conference on World Wide Web Companion*. 759-760.

Biere, S. 2018. "Hate Speech Detection Using Natural Language Processing Techniques", Department of Mathematics, Faculty of Science. Vrije Universiteit Amsterdam.

Brenes Peña, E., Fuentes Rodríguez, C. 2017. "Impoliteness and Thematic Variation in Spanish Television Interviews", *Studies in Media and Communication*, Vol. 5, 1. 50-62.

Brown P., Levinson, S. 1987. *Politeness. Some Universals in Language use*. Cambridge: Cambridge University.

Butler, J. 1997. *Excitable speech: A Politics of the Performative*. Londres: Routledge.

Burnap, P., Williams, M. L. 2016. *Us and them: identifying cyber hate on Twitter across multiple protected characteristics*. EPJ Data Science, 5: 11.

Culpeper, J. 1996. "Toward an Anatomy of Impoliteness", *Journal of Pragmatics* 25. 349-367.

Culpeper, J. 2010. "Conventionalised Impoliteness Formulae", *Journal of Pragmatics* 42. 3232-3245.

Culpeper, J., Haugh, M., Kádár, D. 2017. *The Palgrave Handbook of Linguistic (im)politeness*. Londres: Palgrave Macmillan.

Culpeper, J. 2017. "Impoliteness metalinguistic labels and concepts in

English", *Doing Pragmatics Interculturally. Cognitive, Philosophical and Sociopragmatic Perspectives.* Berlín: Mouton de Gruyter. 135-147.

Douce Munezero, M. 2017. *Leveraging emotion and Word-based features for antisocial behavior detection in user-generated content.* University of Easter Finland, n. 264.

Farzindar, A. 2018. "Natural Language Processing for Social Media", *Synthesis Lectures on Human Language Technologies*, *10*, 2. 1-195. California: Morgan & Claypool Publishers.

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. Faculdade de Engenharia, Universidade do Porto, Oporto.

Gordeev, D. 2016. "Automatic verbal aggression detection for Russian and American imageboards", *Procedia-Social and Behavioral Sciences*, 06/12/2015. Moscú, Rusia.

Herring, S. C. 2004. "Computer-mediated discourse analysis", *Designing for Virtual Communities in the Service of Learning,* S. A. Barab; R. Kling; J. H. Gray (eds.). Nueva York: Cambridge University Press. 338-376.

Miró Llinares, F. 2016. "Taxonomía de la comunicación violenta y el discurso del odio en Internet", *Revista de los Estudios de Derecho y Ciencia Política,* Universitat Oberta de Catalunya. 93-118.

Nahar, V., Li, X., Pang, Ch. 2012. "An Effective Approach for Cyberbullying Detection", *Communication in Information Science and Management Engineering*, *3.* 238-247.

Razavi, A., Inkpen, D.,Uritsky, S., Matwin, S. 2010. "Offensive language detection using multilevel classification", *Advances in Artificial Intelligence*, 6085/2010. 16-27.

Schmidt, A., Wiegand, M. 2017. "A Survey on Hate Speech Detection using Natural Language Processing", *Proceedings of the Fifht International Workshop on Natural Language Processing for Social Media*. Valencia, España, Abril 3-7.

Sperber D., Wilson D. 1986. *La Relevancia. Comunicación y procesos cognitivos*. Madrid: Visor, 1994.

Utsumi, A. 2007. "Verbal Irony As Implicit Display of Ironic Environment:

Distinguishing Ironic Utterances From Nonirony", *Irony in Language and Thought. A cognitive Science Reader,* R. Gibbs; H. Colston (eds.). Nueva York/Londres: Lawrence Erlbaum Associates. 499-529.

Yus, F. 2013. "An inference-centered analysis of jokes", *Irony and Humor. From pragmatics to discourse,* L. Ruiz Gurillo, M. B. Alvarado Ortega (eds.). Ámsterdam/Filadelfia: John Benjamins Publishing Company. 56-82.

Yus, F. 2018. "The interface between pragmatics and internet-mediated communication", *Pragmatics and its Interfaces,* C. Illie, N. R. Norrick. Ámsterdam: John Benjamins. 267-290.

Van Hee, C., Lefever, E., Hoste, V. 2018. "Exploring the fine-grained analysis and automatic detection of irony on Twitter". *Lang Resources & Evaluation.* 52. 707-731.

Waseem, Z., Davison, T., Warmsley, D. Weber, I. 2017. *Understanding Abuse: A Typology of Abusive Language Detection Subtasks,* Proceedings of the First Workshop on Abusive Language Online.

Whitten, S. 2018. "A recognition-sensitive phenomenology of hate speech", *Critical Review of International Social and Political Philosophy.* Belfast, Irlanda.