UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

COLEGIO DE LETRAS HISPÁNICAS

CARACTERIZACIÓN DE LAS OPINIONES MASCULINAS Y FEMENINAS EN UN CORPUS DE *TWITTER*

TESIS

QUE, PARA OBTENER EL TÍTULO DE

LICENCIADA EN LENGUA Y LITERATURAS HISPÁNICAS,

PRESENTA

MADAI RAMÍREZ CISNEROS

ASESOR: MTRO. OCTAVIO AUGUSTO SÁNCHEZ VELÁZQUEZ

CIUDAD UNIVERSITARIA, 2017

CDMX

"Aún soy capaz de grandes cosas, aún soy capaz de sorprenderte a ti, [...], y a todos, aún soy capaz de mover el cielo y la tierra por amor."

Roberto Bolaño, *Amuleto.*

AGRADECIMIENTOS

A la UNAM por todo el conocimiento brindado y a todos los profesores de la FFyL que contribuyeron en mi formación académica. De todos ellos quiero agradecer especialmente a la Dra. María de los Ángeles Adriana Ávila y la Dra. Jeanette Reynoso Noverón por cultivar en mí el amor por la lingüística; a la Dra. Fernanda López Escobedo por introducirme en un área tan interesante como la lingüística forense, y, por supuesto, a la Dra. Ana Aguilar Guevara, mi profesora de seminario de investigación, por creer en cada uno de sus alumnos.

Al proyecto PAPIIT "Enfoques de la semántica distribucional para minería de opiniones" con clave IN403016 por otorgarme una beca para culminar esta tesis.

Al Mtro. Octavio Sánchez, a quien respeto y admiro profundamente, por dirigir mi tesis y por toda su paciencia.

A mis sinodales por sus valiosos comentarios: Dra. María de los Ángeles Adriana Ávila Figueroa, Dra. Fernanda López Escobedo, Mtra. Teresita Adriana Reyes Careaga y Mtra. Bertha Lecumberri Salazar.

Al Dr. Gerardo Sierra Martínez por permitirme formar parte de esta gran familia académica que es el Grupo de Ingeniería Lingüística (GIL) y preocuparse por el bienestar desarrollo de todos sus becarios. A los integrantes del grupo por colaborar con su crítica constructiva a la elaboración de este trabajo.

A Margarita Mota por aparecer en mi vida cuando esta tesis recién germinaba y por acompañarme durante todo el proceso de elaboración. Eres la persona que vivió de cerca mi frustración y desesperación cuando más estancada me sentí. Gracias infinitas por estar ahí.

A Zyanya por ser mi colega de tesis y mi amiga.

A mis amigos de la carrera Marco Antonio Lugo, María del Pilar González Noriega y María del Ángel Montes de Oca Baldespino, por estar conmigo a lo largo de estos cuatro años de licenciatura, por ser mis compañeros de investigación en trabajos finales y por tantos y tantos buenos momentos juntos.

A mis amigas del bachillerato: Ney, Sole, Sue, Roo y Sandy; por permanecer a mi lado a pesar del tiempo y la distancia.

A mi amiga de la infancia, Alisha Morgana Steagall Rivera, por ser la mejor amiga que el mundo pueda pedir y, sobre todo, por existir.

A la persona que se ha convertido en mi mejor amigo y el amor de esta etapa de mi vida, Nico.

Y finalmente, tengo que agradecer a quienes les debo todo lo que he logrado en esta vida: mi familia. A mis padres, Rosalía Cisneros y Agustín Ramírez, por haber cultivado en mí el deseo de superación y perseverancia; por creer siempre en cada una de mis decisiones y apoyarme en todo sentido. A mis hermanos, Amira y Adolfo,

por ser mis modelos y a la vez, mis más grandes admiradores. A mi abuelita, por todo su cariño.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 INTRODUCCIÓN

Internet se ha convertido en una fuente muy útil de información, ya que cada vez resulta más fácil para sus millones de usuarios compartir su propio conocimiento sobre diversos temas. En las redes sociales y los *blogs*[1] millones de personas no sólo comparten información, sino que también expresan su punto de vista sobre distintos hechos, sujetos o entidades; sin embargo, muchas de las opiniones en estos medios permanecen anónimas y otras tantas provienen de perfiles falsos. Conocer quién las escribió puede ayudar a llevar a cabo estudios de mercado para saber en qué grupo posicionar un producto o servicio, pero también en situaciones de peligro cuando se tienen publicaciones de amenaza.

La presente tesis se enmarca en tres áreas de estudio: sociolingüística, lingüística forense y minería de opiniones. La primera es "una ciencia que estudia la lengua [...] en cuanto a instrumento central de comunicación concretamente utilizado en comunidades sociales y que, por tanto, estudia las interrelaciones entre el lenguaje y la sociedad" (Berutto, 1974: 15). Algunos de los temas que se abordan en ella son las variaciones lingüísticas, de manera tanto diacrónica como sincrónica; los comportamientos lingüísticos y los factores sociales que los condicionan (tiempo, espacio, clases sociales y situaciones sociales, principalmente); el cambio lingüístico y las tendencias no estructurales que lo provocan; cómo se relacionan las estructuras

[1] Las redes sociales son sitios de internet donde una persona mantiene contacto con sus distintos grupos sociales: familiares, amigos, compañeros de trabajo o personas con quienes se tiene intereses en común. Las redes más populares son *Facebook*, *Twitter*, *Instagram, Youtube*. Los blogs son páginas de internet donde un individuo publica textos sobre sí mismo a manera de diario o se publican textos sobre un tema particular para difundir y discutir al respecto.

lingüísticas y las estructuras sociales; qué influencia tiene la comunicación lingüística en las interacciones sociales; etc. (Cfr. p. 16).

La segunda especialidad cubre todas las áreas en que la lengua y el derecho se relacionan, es decir, utiliza el conocimiento lingüístico para mejorar la interacción entre los implicados dentro de los procedimientos legales, así como para la resolución de problemas judiciales. Cuando se habla de mejorar la interacción entre los implicados del procedimiento legal, el lingüista forense puede ayudar en la interpretación de una ley y hacer que sea comprensible para los no especialistas; también puede ayudar en la toma de los testimonios de testigos vulnerables, como es el caso de los niños, para que sus palabras no sean tergiversadas o sus derechos violados, o incluso fungir como intérprete en contextos multilingües. La resolución de problemas legales se refiere a realizar un análisis de la prueba lingüística, ya sea de audio o de texto, y emitir un juicio experto. Entre los problemas que la lingüística forense puede ayudar a solucionar se encuentran el plagio y la atribución de autoría. Ante los dos problemas planteados, se considera que las personas tienen un estilo particular para escribir; sin embargo, la edad, escolaridad, lugar de procedencia, estrato socioeconómico e, incluso, el sexo puede influir en las palabras que escogen para expresar sus ideas. Por ello, las preferencias lingüísticas que corresponden a estos grupos han sido estudiadas ampliamente. Con ello, se espera poder perfilar el autor de un texto desconocido, es decir, poder describir las características sociolectales.

La tercera disciplina, minería de opiniones, estudia la extracción automática de información subjetiva (como las creencias, los sentimientos, las desvalorizaciones, etc.), la cual pretende encontrar, extraer y clasificar dicha información. Una de sus aplicaciones está en el ámbito empresarial, pues permite conocer qué opinan los

usuarios sobre un producto o servicio. Sin embargo, también puede utilizarse para el análisis de textos literarios (por ejemplo: *¿sobre qué sentimientos se habla en la obra de un autor?*) y análisis del discurso (por ejemplo: *¿cómo se construye la imagen de un personaje público en la prensa?*).

Con la presente tesis, se realizó un análisis sociolingüístico basado en minería de opiniones, puesto que se hizo un análisis lingüístico de las opiniones en relación con el factor social género y se espera que dicho análisis pueda ser aplicado para el perfilamiento de sexo. Se debe realizar un estudio granular sobre cómo opinan hombres y mujeres para observar las diferentes estrategias que cada sexo utiliza para expresar su subjetividad y así, evitar afirmaciones sobre si el estilo femenino es más "emotivo" o "personal" y el masculino más "informativo" a partir géneros discursivos, como las anécdotas o estados de *Facebook*, considerados como más subjetivos (Johnstone, 1993; Schwartz, 2013). Por lo tanto, aquí se consideró a la opinión una unidad subjetiva mínima discursiva.

La metodología fue la siguiente: primero, se elaboró un corpus de *tweets* donde se controló la variante dialectal, el nivel de escolaridad y el grupo etario; segundo, se llevó a cabo la clasificación de *tweets* en "opiniones" y "no opiniones" de acuerdo con los elementos que debe tener la opinión explicados en el capítulo minería de opiniones; tercero, se clasificará el elemento subjetivo de acuerdo con la categoría léxica o tipo de frase (locuciones y enunciados fraseológicos) a la que pertenece; cuarto, se procedió a realizar un conteo de frecuencias sobre los elementos encontrados, y finalmente, se realizó un análisis estadístico para ver si los resultados obtenidos son dependientes. Es decir, si los resultados pueden repetirse en otro corpus de condiciones similares.

1.1 Motivación

Existen algunos prejuicios lingüísticos sobre cómo hablan los hombres y las mujeres. Popularmente se cree que las mujeres son más corteses, menos groseras y que un hombre nunca utilizaría palabras como 'lindo' o 'precioso' para describir un objeto, para muestra Lakoff (1973) afirmó que este tipo de adjetivos eran exclusivos de las féminas; sin embargo, se necesitan más estudios cuantitativos para afirmar o refutar los juicios *a priori*. La presente investigación parte de la idea de que hombres y mujeres hablan diferente y por ello se decide comprobarlo dentro de un corpus que se asemeja al comportamiento "real" del hablante, enfocándose en las opiniones, en específico, con el propósito de distinguir qué elementos lingüísticos prefieren utilizar para emitir una valoración.

Este trabajo surgió dentro de un proyecto CONACYT del Grupo de Ingeniería Lingüística (GIL) denominado "Caracterización de huellas textuales para análisis forense" con clave 215179, en el cual se buscaba encontrar los elementos lingüísticos que permiten distinguir depredadores sexuales, quienes utilizan internet para encontrar posibles víctimas. Se ha denominado *huellas textuales*, a las características de estilo, como de producción lingüística que permiten caracterizar a un individuo dentro de un grupo específico, por ejemplo: el sexo, grupo etario o la escolaridad del autor. Con este proyecto se pretendía realizar análisis estadísticos y de lingüística computacional, y así darle mayores herramientas al perito lingüista para perfilar un autor.

Actualmente, se está desarrollando el proyecto PAPIIT "Enfoques de la semántica distribucional para minería de opiniones" con número IN403016 cuyo objetivo es encontrar una metodología que permita caracterizar, identificar, detectar

y extraer sentimientos. A partir de este proyecto, surgió la idea de tomar como marcadores distintivos de sexo la forma de emitir opiniones.

1.2 Justificación y planteamiento del problema

Existe poca investigación sobre la diferenciación lingüística de género (masculino-femenino) en español. Cabe destacar que en la UNAM existe una tesis denominada *Sociolingüística de los adjetivos calificativos en un corpus del español mexicano* (Rivera; 2015), la cual trata de ver las preferencias de estos adjetivos para realizar una opinión; sin embargo, dichos adjetivos no son los únicos elementos lingüísticos que se pueden utilizar para emitir una valoración; por ejemplo: algunas clases de verbos, adverbios, sustantivos, interjecciones, locuciones y enunciados fraseológicos también se pueden utilizar para tal propósito. Además, la disertación mencionada tiene el propósito de ver el cambio diacrónico sobre el uso de dichos adjetivos, para ello utiliza un corpus de entrevistas de 1970 y otro del 2000. En este caso se utilizará un corpus de mayor actualidad (entre otras características) (*Twitter*), de modo que se espera encontrar cosas diferentes, incluso sobre el uso de dichos adjetivos (adjetivos nuevos, préstamos, etc.). Además, no siempre que se usen adjetivos evaluativos se está emitiendo una opinión.

Por lo tanto, se propone la siguiente pregunta de investigación que motiva la presente tesis:

¿Cuáles son las diferencias lingüísticas existentes entre hombres y mujeres para expresar una opinión?

1.3 Hipótesis

Frente al problema planteado se tiene la siguiente hipótesis:

Cada sexo tiene preferencias sobre una clase de palabra o frase para dar una valoración.

1.4 Objetivos

Con el propósito de satisfacer el problema planteado en 1.3 se propone lo siguiente como objetivo principal.

1.4.1 Objetivo general

Encontrar qué clase de palabras o frases prefieren utilizar hombres y mujeres con escolaridad superior para emitir una opinión dentro de una red social.

Para conseguir este objetivo se propusieron los sucesivos objetivos particulares:

1.4.2 Objetivos particulares

- Elaborar un corpus de *Twitter*.
- Identificar las opiniones dentro de los *tweets*.
- Clasificar el enunciado de opinión de acuerdo con la clase de palabra o frase a la que corresponde.
- Elaborar una prueba estadística para aceptar o rechazar dependencia.

1.5 Descripción de la estructura de la tesis

A continuación, se describirá la estructura de esta tesis.

La presente introducción constituye el primer capítulo.

En el segundo capítulo, se dará la definición de lingüística forense, así como la explicación de algunas de las tareas que realiza, entre ellas se encuentra el perfilamiento de autor, tarea para la cual se pretende contribuir con este trabajo.

En tercer capítulo, se definirá el área de minería de opiniones y se mencionarán algunas de sus aplicaciones. Se hablará sobre la distinción entre lo que es objetivo y lo que es subjetivo, qué son las opiniones y cuáles son sus tipos, así también, en ese capítulo se dará un breve estado de la técnica sobre la clasificación de opiniones en *Twitter*: el uso de diccionarios de sentimientos como una herramienta para encontrar polaridad en esta red.

En el capítulo cuatro, se describirá el estado del arte sobre las preferencias lingüísticas de cada sexo; primero, se planteará la problemática de la subjetividad como inherente a un sexo vs las exigencias del género discursivo que se utiliza; segundo, se abordarán los trabajos relacionados con las preferencias lingüísticas en los medios de comunicación social tales como *blogs* y redes sociales, y finalmente, se explicarán las investigaciones que utilizan minería de opiniones para el perfilamiento automático.

En el capítulo cinco se expondrán las categorías léxicas que expresan de alguna manera una valoración sobre una entidad. Se describirán los sustantivos, adjetivos, adverbios, interjecciones, locuciones y enunciados fraseológicos: así como las clasificaciones semánticas de algunas de estas categorías que permitieron considerar los términos como valorativos.

En el capítulo seis se describirá detalladamente la conformación del corpus de *Twitter*, así como las características más importantes de esta red social. Posteriormente, se explicarán los criterios tomados para el etiquetado del corpus, los problemas encontrados y cómo fueron resueltos.

En el capítulo siete, se mostrarán los resultados obtenidos: qué sexo opina más en *Twitter* y sobre qué, si son más positivos o negativos en sus opiniones, qué tipo de opinión prefieren utilizar, y por supuesto, el objetivo de esta tesis, qué tipo de unidad léxica o frase prefieren utilizar para opinar.

En el octavo y último capítulo se dirán las conclusiones obtenidas en este trabajo, se dará un resumen de cada capítulo y se expondrá el trabajo que se desea desarrollar en el futuro.

Por último, cabe aclarar que a lo largo de esta tesis se prefirió utilizar "sexo" en lugar de "género" para referirse a hombres y mujeres, debido a las confusiones que provoca con "género discursivo" y "género gramatical". Además, se sabe que existen diferencias entre los primeros dos términos: por un lado, el género tiene que ver con cómo se identifica a sí mismo el individuo, si como hombre, mujer o posee una identidad fuera de estas dos categorías; por otro, el sexo incluye el tipo de cuerpo, los cromosomas, las hormonas y los órganos reproductivos (Nagoshi, 2014). En este trabajo no se tomó en cuenta si las preferencias sexuales de los participantes o cómo se asumían a sí mismos condicionan sus preferencias lingüísticas, eso forma parte del trabajo futuro; por lo tanto, el término más adecuado era "sexo".

2 LINGÜÍSTICA FORENSE

Este capítulo tiene por objetivo detallar una de las disciplinas en las que se enmarca esta tesis: lingüística forense. En primer lugar, se definirá el área de especialidad; en segundo se mencionarán las áreas de aplicación; en tercero, se describirá con mayor detalle la tarea para la cual se pretende contribuir con esta investigación: perfilamiento de autor.

2.1 Definición

La Asociación Internacional de Lingüística Forense (IAFL, 2013) define esta disciplina como "la interfaz entre la lengua y el derecho". Esta aseveración parece demasiado extensa y vaga, pero justamente "cubre todas las áreas en las que la lengua y el derecho se interrelacionan" (Turell, 2005:13), ya sea en la redacción de leyes, la comunicación en el juzgado o, incluso, los problemas legales que requieren el análisis de un experto de la lengua. Esta es la definición amplia del área; sin embargo, existe una más acotada en la que algunos especialistas consideran que la lingüística forense se encarga únicamente del análisis de pruebas lingüísticas, textuales u orales, para la resolución de casos legales. Al respecto Turell dice que "se refiere especialmente a la utilización de pruebas lingüísticas en los juicios, y por tanto, a la actuación de los lingüistas en contextos jurídicos y judiciales" (Turell, 2005:13). Ahora bien, entiéndase como prueba lingüística cualquier texto oral u escrito que se relacione de alguna manera con un contexto criminal o legal; por ejemplo: una nota de rescate, de amenaza, una grabación de extorsión, etcétera (cfr. Olson, 2013: 1). Otros ejemplos que pueden citarse incluyen comunicaciones personales por vía electrónica (*Twitter*, *Facebook*, etc.), bitácoras o diarios personales (*blogs*, *vlogs*, etc.)

Analizar pruebas lingüísticas requiere del conocimiento de muchos y diversos métodos y técnicas de las distintas subáreas de la lingüística. Como ejemplificación, tómese en cuenta que para determinar la intencionalidad criminal se hace un análisis del discurso sobre la declaración del implicado, mientras que para atribución de autoría se utilizan "los métodos y técnicas de la sociolingüística y la dialectología" (Shuy, 2005: 21). Sobre estas dos tareas se hablará en los siguientes apartados.

2.2 Áreas de aplicación

En el apartado anterior se habló de dos definiciones sobre la ciencia que nombra este capítulo: una amplia y otra restrictiva. Aquí se expondrán algunas de las aplicaciones de dicha rama del saber, pero desde su definición amplia.

Existen tres formas en las que el lenguaje y la ley se interrelacionan: el **lenguaje de la ley**, el **lenguaje del proceso judicial** y el **lenguaje como evidencia**. En la primera, los lingüistas ayudan en el análisis, comprensión e interpretación de documentos legales (Olson, 2013:10), debido a que el lenguaje de la ley suele ser ambiguo y lejano para muchos de los procesados. Además, se piensa en las minorías lingüísticas y en cómo la ley aborda sus derechos (IAFL, 2013).

La segunda forma, el lenguaje del proceso judicial, tiene que ver con la interacción entre los diversos implicados en un crimen a lo largo del procedimiento judicial, tanto en las declaraciones, como en la sala de justicia. Por lo cual, se analizan las entrevistas con niños y testigos vulnerables, el lenguaje de la policía, jueces y jurado. Asimismo, se contribuye a que los implicados interactúen en el juzgado, ya sea en la traducción en contexto multilingües o, en la simple interpretación del lenguaje técnico de los abogados y los jueces (IAFL, 2013).

En la tercera manera, el lenguaje puede ser utilizado como evidencia cuando se tienen pruebas lingüísticas. Si las pruebas son textuales, se pueden realizar las siguientes tareas: identificación de plagio (se determina si una obra es la copia intelectual de otra), atribución de autoría (se determina quién es el autor de un texto, para ello se compara el estilo de varios autores en distintos textos y se identifica quién tiene un estilo más parecido al del texto en disputa), perfilamiento de autor (cuando se desconoce quién pudo haber escrito un texto, se describen las características socio-culturales de un posible autor a partir del escrito). Si las pruebas son grabaciones de voz, un testigo auditivo se encarga de identificar al autor del delito, o si las pruebas contienen demasiado ruido, e lingüista trata de identificar el contenido de las grabaciones (cfr. Olson, 2013: 11).

2.3 Perfilamiento de autor

Existen algunos casos en los que se desconoce el autor de un texto y no existe un grupo cerrado de posibles sospechosos (por ejemplo: casos de terrorismo, correo de odio y de acecho), entonces el lingüista forense debe reconocer dentro de los textos la presencia de características del lenguaje que ayuden a identificar los antecedentes regionales del escritor, su edad, el sexo, el nivel educativo, la raza, la etnicidad, la ideología política, la profesión, e incluso, la religión como pistas de autoría (Shuy, 2014: 73; Shuy, 2005: 19). A esto se le conoce como perfilamiento de autor. Cabe aclarar que el perfilamiento lingüístico no proporciona una identificación exacta del sospechoso, en todo caso, dice a la policía qué características tiene la persona que deben buscar.

Otra definición de la misma actividad fue explicada por Campbell and DeNevi como:

> it describes how the suspect's language matches social, economic, education level, and other information that previous sociolinguistic research has identified to be characteristic of specific societal groups. Its purpose is to help law enforcement narrow down its list of existing suspects or to suggest directions for locating new ones (2004 citado en Shuy 2014).[2]

Los estudios previos en sociolingüística han permitido identificar algunas de las características estilísticas que comparten varios grupos sociales; pues la preocupación de esta disciplina es entender y explicar "el hecho lingüístico en relación con el grupo social o el grupo de individuos que lo utiliza" (Areiza, 2012: 6). Por tanto, los factores dialectales y sociales mencionados con anterioridad influyen en el estilo de un autor.

El idiolecto es la manera en que cada individuo se expresa al hablar. Cuando intenta plasmar sus ideas en papel elige repetitiva e inconscientemente una forma lingüística sobre otras, lo cual conforma su estilo (cfr. Mcmenanin, 2010: 488). Cuando se realiza atribución de autoría, se trata de encontrar las preferencias lingüísticas que lo hacen diferente de los demás. En cambio, cuando se realiza perfilamiento, se trata de encontrar las características compartidas con otros grupos sociales.

Es factible realizar perfilamiento de autor con procedimientos automáticos si se utilizan algoritmos de aprendizaje de máquina. En el estado del arte se presentarán algunos de los trabajos que han realizado esta tarea utilizando diversas características predictivas y se han probado en corpus diferentes.

---

[2] "Esta describe cómo el lenguaje del sospechoso encaja con información social, económica, nivel educativo y de otro tipo que los estudios sociolingüísticos previos han identificado como característicos de grupos sociales específicos. Su propósito es ayudar a la aplicación de la ley para reducir su lista de sospechosos existentes o sugerir direcciones para ubicar nuevos." (La traducción es mía).

3 MINERÍA DE OPINIONES

En esta investigación se aplicó un análisis con base en los métodos y definiciones de minería de opiniones, por lo que es importante dedicar un capítulo completo para definir el área, así como mencionar algunas de sus aplicaciones. Sin embargo, antes de pasar de lleno a aclarar de qué se trata minería de opiniones, tenemos que explicar qué es subjetividad, puesto que esta es su objeto de estudio. Cuando ya se haya entrado en materia, se dirá qué es una opinión, cuáles son los tipos de opiniones y sus características, elementos clave para el desarrollo de la metodología. Finalmente, se consideró necesario incluir los diccionarios de sentimientos, herramientas para la extracción automática de subjetividad, así como algunos de los trabajos en los que se haya realizado clasificación automática de opiniones en *Twitter*.

3.1 Subjetividad

De acuerdo con Kerbrat (1986) la diferencia entre la información objetiva y subjetiva tiene que ver con la manera como el sujeto enunciador verbaliza un objeto referencial, real o imaginario. Para ello, el emisor selecciona ciertas unidades del repertorio léxico y sintáctico de su lengua y las presenta de dos formas:

1. con un **discurso objetivo.** Se refiere a hechos acerca del enunciador en relación con el mundo, es decir, es verificable en mayor o menor medida (Ortony, Clore y Foss, 1987: 343; Liu, 2012: 27) y el enunciador trata de borrar todo rastro de su existencia como enunciador individual (Kerbrat, 1986). Por ejemplo: "*El iPhone es un producto Apple.*"
2. con un **discurso subjetivo**. Se refiere a estados privados donde el enunciador se muestra explícitamente ("lo encuentro feo") o se reconoce implícitamente ("es feo") como la fuente evaluativa de la afirmación (cfr. Kerbrat, 1986: 93).

Los discursos subjetivos pueden verse como la expresión de un estado privado. Estos son estados que no están abiertos a observación o verificación: se puede observar que una persona afirma que Dios existe, pero no que cree que Dios existe. Por lo tanto, las creencias, los sentimientos, los acuerdos, desacuerdos y las opiniones son estados privados (Quirk *et al.*, 1985).

Cabe añadir que los investigadores Traugott y Langacker, de manera independiente, han elaborado trabajos sobre la subjetividad. Por un lado, Traugott (1989) apunta que la subjetificación es una de las tendencias de cambio semántico, en la que el significado se vuelve más subjetivo, es decir, tiende a un incremento de la codificación de la actitud del hablante, ya sea de la creencia, la evaluación de la verdad o el compromiso de aserción. Existe graduación dentro de la subjetividad, donde una enunciación es más subjetiva cuando incluye al emisor en la comunicación (***yo*** *concluyo*, ***yo*** *pienso*). Por otro, Langacker (1985) describe a la subjetividad como un fenómeno gradual que tiene que ver con la forma en cómo se construye una expresión. Todas las situaciones son construidas por los participantes del acto de habla, pero algunas son expresadas de una manera objetiva, en el sentido en que los eventos y los participantes del mismo están sobre el escenario como focos de atención mientras que el acto de habla en sí mismo y sus participantes están fuera del escenario, sin perfilar como espectadores. En una construcción subjetiva, las situaciones están expresadas de una manera que alinean algunos aspectos del evento y sus implicados con el acto de habla y los participantes de la comunicación, y por lo tanto, con la vista de la escena. Un caso sería cuando el hablante se ubica como uno de los participantes del evento, y si lo hace y no es de ninguna manera un participante de la escena, se dice que la expresión tiene el grado más alto de subjetividad.

3.2 Definición de minería de opiniones y sus aplicaciones

A través de internet se comparte diariamente una gran cantidad de información de manera fácil y rápida, y de la misma forma se tiene acceso a ella. Dicha información puede ser sobre un acontecimiento que se originó en un lugar distinto al que se encuentra quien comparte los datos, o monografías y avances sobre distintas ciencias, esto es información objetiva, o también, detalles sobre el día a día de una persona. La gente publica en sus redes sociales (como *Facebook* o *Twitter*) qué comió, qué película vio o qué hizo durante el día. Para muchos, esto puede parecer banal; sin embargo, para otros puede resultar muy valioso puesto que en dichas redes no sólo se comparten fotos, también se comparten nuestras valoraciones sobre diversos productos, servicios, política o diversos temas de actualidad, así como deseos, anhelos, esperanzas, etcétera, esto es información subjetiva. Existen, incluso, páginas que se dedican exclusivamente a la creación de reseñas sobre temas, lugares u objetos particulares como películas, hoteles y restaurantes. Los millones de usuarios de internet pueden consultar dichas páginas para saber qué dijo alguien más sobre un lugar que le interesa visitar y, con base en la experiencia de otro usuario, tomar una decisión. A la disciplina que se encarga de extraer automáticamente información subjetiva se le conoce como minería de opiniones.

Minería de opiniones (MO), o también conocida como análisis de sentimientos, es una reciente subdisciplina entre la recuperación de información y la lingüística computacional, la cual no se preocupa únicamente del tema sino de las opiniones que se expresan (Esuli, 2006: 417). En su libro *Sentiment Analysis and Opinion Mining,* Bing Liu la define como: el campo de estudio que analiza las opiniones de las personas, los sentimientos, las evaluaciones, estimaciones, las actitudes y las emociones hacia entidades como productos, servicios, organizaciones, individuos,

problemas, eventos, temas y sus atributos (2012: 7). Por su parte, Bora (2012: 42) lo trata como: el estudio de determinar computacionalmente si dado un fragmento de texto, este es indicativo de sentimientos positivos o negativos. En resumen, esta disciplina trata de extraer computacionalmente la información subjetiva de un texto (esta puede ser un sentimiento, una valoración, una creencia), así como cuál es el tema y si la opinión es positiva o negativa.

Existen tres subtareas en MO: 1) discriminar si la información es subjetiva u objetiva; 2) en caso de que la información sea subjetiva, debe identificarse la polaridad, es decir, si esta es negativa o positiva, y 3) poder graduar la información subjetiva en baja, media o altamente positiva o negativa (cfr. Esuli, 2006:417). En esta tesis, primero se realizó la tarea de discriminar entre información objetiva y subjetiva, puesto que primero se separaron las opiniones de los hechos o descripciones objetivas, y después se realizó una discriminación sobre la polaridad de las opiniones.

Ya que se definió la disciplina, resulta pertinente hablar de sus múltiples aplicaciones. Se puede analizar una obra literaria para saber sobre qué emoción se prefiere hablar en ella. En el ámbito empresarial se puede investigar qué opinan las personas sobre un producto o servicio para mejorar su calidad. En la política, para saber la aceptación de un político y adecuar su imagen. Las distintas redes sociales representan una ventaja para empresas y campañas políticas debido a que, realizar un análisis de MO en ellas resulta más económico que hacer las tradicionales encuestas y se cubre un porcentaje mayor de la población. Cabe añadir que se puede buscar información subjetiva en textos donde se pretende ser objetivos, como las noticias, ya que algunos periodistas o periódicos adoptan una postura frente a un hecho.

3.3 Opiniones

De acuerdo con Liu (2012: 26), existen dos tipos de opiniones: las opiniones explícitas y las opiniones implícitas. Las opiniones explícitas son declaraciones subjetivas que expresan una opinión regular o comparativa; mientras que las implícitas son declaraciones objetivas que implican una opinión comparativa o regular. El segundo tipo de opiniones están fuera del alcance de la investigación.

De las formas explícitas en las que se expresa una opinión, se pueden distinguir diferentes tipos. Principalmente dos: una en la que se expresa una opinión de manera directa sobre algo (*Me gustan los días lluviosos*) y otra en la que no se puede decir si algo es por sí solo positivo o negativo, sino que se enmarca en una comparación con respecto a otra cosa (*Carol es más bonita que Elena*). A continuación, explicamos a detalle cada una de estas.

3.3.1 Opiniones regulares

Una opinión es la valoración positiva o negativa que alguien hace sobre un hecho o entidad y posee los siguientes elementos: el objeto de opinión, el sentimiento (o polaridad), opinador. El objeto de opinión puede ser cualquier producto, servicio, tema, persona, organización, evento, problema.

Una de las formas para anotar una opinión se encuentra en Liu (2012: 18), quien la define con 4 elementos:

1. **g**: el objeto del que se opina.
2. **s** (*sentiment*): el sentimiento que se tiene sobre el objetivo u objeto del que se opina.
3. **h** (*opinion holder*): el opinador.
4. **t** (*time*): cuando fue expresada la opinión.

Otra propuesta de anotación, y a la cual se ajustó el etiquetado de opiniones regulares de este trabajo, fue la hecha por Wiebe *et al.* (2005: 6-7), quienes mencionan que la opinión contiene los siguientes elementos:

1. **Ancla de texto:** un punto en el fragmento de texto que representa el evento de habla o la mención explícita del estado privado.
2. **Fuente:** la persona o la entidad que expresa el estado privado.
3. **Objeto:** el objeto o el tema del estado privado, por ejemplo: de qué trata el acto de habla o el estado privado.

Además de esos elementos, las opiniones contienen las siguientes propiedades:

1. **Intensidad:** la intensidad del estado privado (baja, media, alta o extrema).
2. **Tipo de actitud:** este atributo representa la polaridad del estado privado. Los valores son positivo, negativo, otro, o ninguno**.**

La propuesta de Liu contiene el elemento *tiempo*, este es importante cuando se requiere darle seguimiento a la opinión pública sobre un acontecimiento, ver cómo cambia en un periodo o cuando deja de hablarse de él. A diferencia de ella, la propuesta de Wiebe *et al.* no sólo está pensada para opiniones, sino también para cualquier tipo de estado privado que pueda encontrarse expresado en un texto. En ella se realiza una distinción de tres elementos:

1) la **expresión explícita de un estado privado**, por ejemplo: "Los Estados Unidos *temen* un derrame",

2) Los **elementos expresivos subjetivos**, por ejemplo: "El reporte estaba *lleno de absurdos*", y

3) Un **evento de habla que expresa un estado privado,** por ejemplo**:** **"**El reporte estaba lleno de absurdos" *dijo* Xirao Nima (Wiebe, 2005: 5).

En 1) se marca un verbo que expresa un estado privado; en 2), una frase que expresa una valoración, y en 3), un verbo que denota un acto de habla, el cual introduce un estado privado. Distinguir entre estos elementos permite determinar si la valoración predomina sobre todo un documento y no sólo sobre una frase, debido

a que se marca cuando la opinión pertenece al autor del documento y cuando a otro; por ejemplo, en las noticias ayuda a ver la postura del reportero con respecto al tema pues se distingue el tipo de verbo de habla (si es neutral, negativo o positivo).

Se debe aclarar que algunas opiniones pueden ser emocionales (1) pero no todos los enunciados emocionales expresan opiniones (2):

(1) Estoy feliz con este carro.
(2) Estoy muy triste hoy.

3.3.2 Opiniones comparativas

Las sentencias comparativas son una oración que expresa una relación basada en similitudes y diferencias de más de un objeto. Un objeto es una entidad que puede ser una persona, producto, una acción, etc. (cfr. Jindal and Liu, 2006: 4). Es decir, dichas opiniones cumplen con los siguientes elementos:

1. **E1**: entidad 1.
2. **E2**: entidad 2.
3. **A**: aspectos en los que se comparan las entidades 1 y 2.
4. **PE**: entidad que prefiere el opinador.
5. **h**: opinador.
6. **t**: el tiempo en el que se expresa la opinión (cfr. Liu, 2012:112).

Como puede observarse en el esquema anterior, no es necesario incluir la polaridad de la opinión puesto que este tipo de opiniones no expresan directamente una opinión positiva o negativa. En su lugar, se comparan las entidades ordenándolas por los aspectos que comparten. Esto es, se expresa una preferencia por alguna o algunas entidades. Para propósitos de aplicación uno puede asignar una opinión positiva por la entidad que prefiere el opinador (cfr. Liu, 2012: 115).

Existen dos clases de comparaciones: las graduables y no graduables. Las primeras, a su vez se subclasifican en:

1. Graduables no iguales: relaciones del tipo *mayor o menor que* que expresan un orden de algunos objetos que consideran ciertas características. Este tipo incluye preferencias de uso.
2. Igualatorio: relaciones del tipo *igual a* o *tanto como* que fijan dos objetos como iguales con respecto a algunas características.
3. Superlativo: oraciones del tipo *el mejor o el peor*, las cuales gradúan un objeto sobre todos los otros (cfr. Jindal and Liu, 2006: 4).

Las no graduables son oraciones que comparan características de dos o más objetos, pero no los gradúan, por lo cual, aunque pueden expresar una opinión sobre una entidad esta puede ser difícil de verse:

1. Entidad A es similar a o diferente de B basado en los aspectos que comparten. Ejemplo: *Coca sabe diferente a Pepsi.*
2. Entidad A tiene el aspecto a1 y entidad B tiene el aspecto a2. (Los aspectos son usualmente intercambiables). Ejemplo: *Las computadoras de escritorio usan bocinas externas mientras las laptops usan bocinas internas.*
3. Entidad A tiene aspecto a pero B no lo tiene. Ejemplo: *Los teléfonos Nokia vienen con auriculares pero los Iphone no.* (cfr. Liu, 2012: 111).

Una forma de extraer opiniones comparativas es utilizando los adjetivos y adverbios comparativos; sin embargo, estas palabras no son la única forma de comparar dos objetos (también se pueden utilizar palabras como *preferir*), y en algunos casos, aunque se tengan palabras comparativas tampoco se está realizando una comparación (cfr. Jindal *et al.*, 2006); por ejemplo: "No puedo estar *más* de acuerdo contigo".

Si bien las palabras o adverbios comparativos no siempre ayudan a identificar opiniones comparativas, estos permiten localizarlas de manera más rápida; por lo que

se han desarrollado lexicones que contengan estos y otros tipos de palabras que permiten extraer de manera automática los dos tipos de opiniones. A continuación se hablará más a detalle sobre ellos.

3.4 Diccionarios de sentimientos

Todas las palabras pueden conllevar un significado connotativo afectivo, incluso aquellas que aparentemente *no lo son* pueden evocar experiencias placenteras o dolorosas debido a su relación semántica con conceptos o categorías emocionales. El poder afectivo de unas palabras es parte del imaginario colectivo (*guerra*, *mamá*, *fantasma*) y otras refieren directamente a estados emocionales (*miedo*, *alegría*) (cfr. Strapparava y Mihalcea, 2007). Las palabras que son utilizadas comúnmente para expresar sentimientos positivos (*bueno*, *maravilloso* y *sorprendente*) y negativos (*terrible* y *malo*) han sido recogidas dentro de diccionarios con el fin de proporcionar una herramienta que permita la identificación automática de sentimientos y opiniones. A estas herramientas se les conoce como diccionarios de sentimientos.

Estos lexicones están anotados dependiendo de: a) la polaridad a la que pertenece la palabra, es decir, si esta es positiva, negativa o neutra y le proporcionan un valor numérico para expresar la intensidad de la misma, qué tan positiva o negativa es, o qué tan seguros están los anotadores de ello, y otros diccionarios, de acuerdo con b) el sentimiento que expresa la palabra, esto es, si está relacionada con el miedo, la tristeza, la felicidad, etc. (Rangel *et al.*, 2014). Así, contando si las palabras que tiene un texto aparecen en el diccionario, podría decirse con ayuda de un algoritmo si este es más positivo o negativo, o hacia qué sentimiento se orienta el texto en su mayoría.

Existen numerosas herramientas de este tipo para el inglés, como el *ANEW* (*Affective Norms for English Words*) (Bradley y Lang, 1999), el *SentiWordNet* (Esuli y Sebastiani, 2006) y *WordNet-Affect* (Strapparava y Valitutti, 2004). Estos dos últimos fueron desarrollados a partir del *WordNet* (diccionario de la lengua inglesa que contiene la relación semántica de sus términos, ya sea hiponimia, hiperonimia, antonimia, sinonimia (Miller, 1995)).

SentiWordNet es un diccionario del tipo a mencionado arriba en el que todas las palabras están relacionadas con tres calificaciones numéricas: objetiva, positiva y negativa; puesto que los autores asumieron que un mismo *synset* (grupos de sinónimos que contienen indicadores que describen la relación semántica entre un *sysnet* y otro) puede adoptar cada una de las propiedades dependiendo del contexto. Cada una de las tres calificaciones fue valorada del 0.0 al 1.0, para cada término la suma es igual a 1.0. Esta valoración está relacionada con la confianza que se tiene en la etiqueta de un término y no en qué tan fuerte el término parece poseer dicha propiedad.

Fueron seleccionados 14 *synsets* paradigmáticamente positivos y negativos (ejemplo: asqueroso = negativo, bonito = positivo), a partir de ellos fueron adheridos automáticamente todos los synsets que están conectados por relaciones léxicas de *WordNet*, aquellos conectados por relaciones de sinonimia adoptaron la misma polaridad y a aquellos de antonimia, les asignaron la relación opuesta. Posteriormente, cada *synset* obtuvo una representación vectorial y estas fueron introducidas en un clasificador supervisado, el cual genera dos clasificadores binarios: 1) positivo y no positivo, frente 2) negativo y su categoría complementaria, no negativo. Son positivos los términos que fueron clasificados como positivos y no negativos, y negativos los que fueron clasificados como negativos y no positivos. Los

términos neutros fueron clasificados por los clasificadores a veces como positivos y otras como negativos.

Por otro lado, *WordNet-Affect* es un diccionario del tipo b. Primero, el recurso llamado *AFFECT* fue realizado manualmente, este es una base de datos que contiene 1309 palabras que están relacionadas directa o indirectamente con estados mentales (539 sustantivos, 517 adjetivos, 238 verbos y 15 adverbios). A cada caso le fue asignado información léxica y la categoría afectiva. La información léxica contiene la correlación entre términos italianos e ingleses, partes de la oración, sinónimos, antónimos, definición. Con ayuda del *AFFECT*, fue seleccionado automáticamente un grupo de *synsets* de *WordNet* adecuados para representar los conceptos afectivos, a cada uno de ellos le asignaron una o más etiquetas afectivas (estado anímico, respuestas emocionales, situaciones que incitan a las emociones) para contribuir a precisar el significado afectivo. Fue asumido que el sentido afectivo se conserva en relaciones de sinonimia, antonimia, derivación, pertenencia, atribución y fueron adheridos los *synsets* que estaban relacionados en *WordNet*. Manualmente, fue revisado que los *synset* que tuvieran relaciones de hiperonimia efectivamente poseyeran información afectiva. Actualmente, *WordNet-Affect* contiene 2,874 *synsets* y 4,787 palabras.

No obstante, para la lengua española se encontraron pocos diccionarios como el *SentiText*, la traducción del *ANEW* al español y el elaborado por el IPN. Los dos primeros son del tipo a y el tercero es del tipo b.

El *SentiText* (Ortíz *et al.* 2010) contiene palabras que fueron extraídas del diccionario de sinónimos del español *Open Office*. A esas palabras les asignaron una valencia de -2, -1, 0, 1, 2, donde -2 es muy negativo y 2 muy positivo. El lexicón también posee locuciones y expresiones multipalabra.

Redondo *et al.* (2007) tradujeron el *ANEW* al español con ayuda de un filólogo especialista del inglés. Enseguida, 680 participantes (560 mujeres, 120 hombres) de 18 a 25 años debían calificar con escalas de 1 a 9, valencia (agradable a no agradable), excitación (calma a excitación), y dominio (va de control a fuera de control).

Para el recurso elaborado por el IPN (Rangel *et al.,* 2014), los autores tradujeron las palabras del inglés del *WordNetAffect* de manera automática con el traductor de *Google*, traductor *Babylon* y el *English-Spanish Interpreter Pro.* Luego, integraron las traducciones de todas las palabras y eliminaron las repeticiones. Después, verificaron su existencia con el diccionario de María Moliner (1996). Posteriormente, las palabras fueron clasificadas en seis sentimientos: alegría, repulsión, miedo, enojo, tristeza y sorpresa. El siguiente paso fue evaluar cada término: 108 personas debían calificar qué tan relacionado estaba el vocablo con la emoción (nula, baja, media, alta). Únicamente fueron promediadas las valoraciones de las personas que tenían un alto nivel de concordancia. Como dato extra, este recurso es pequeño y no contiene locuciones.

Aunque todos estos recursos son útiles, no agotan el problema. Dos de ellos no poseen locuciones (el recurso del IPN y la traducción del *ANEW*), ni todas las palabras que pueden tener un sentido afectivo, y también debe considerarse que no todas estas palabras funcionan en todos los contextos. En los siguientes ejemplos extraídos de Liu (2012) puede notarse que, aunque las palabras en cursivas son palabras que denotan sentimientos, el enunciado no refleja una polaridad sobre una entidad en particular y por tanto no es una opinión.

Hay casos donde se realiza una pregunta:

(3) ¿Alguien puede decirme dónde encontrar un *buen* teléfono?

Y otros donde se expresa una condición:

(4) Si alguien hiciera un carro *confiable*, lo compraría.

3.5 Clasificación de sentimientos en *tweets*

Las opiniones se pueden encontrar en diversas aplicaciones y redes sociales. No obstante, *Twitter* ha atraído recientemente la atención de MO por el hecho de que en ella se comparte al instante, es decir, en el momento en el que sucede un acontecimiento, temas de interés general. A continuación se describen algunos de los trabajos realizados para la clasificación automática de sentimientos en *Twitter*, en ellos se utilizan métodos y herramientas de aprendizaje de máquina; sin embargo, también pueden utilizarse lexicones, como se verá en el estado del arte sobre perfilamiento automático en el capítulo 4.

Uno de los primeros documentos sobre la detección de polaridad en *Twitter* fue el de Go *et al.,* (2009). Utilizaron 3 métodos de *Machine Learning*: *Naive Bayes*, Máxima Entropía y Máquina de Vectores Soporte (SVM) (cfr. Manning, 1999: 589, 539, 43), y lo compararon con una referencia de palabras clave. Reportaron que los algoritmos de aprendizaje de máquina se desarrollan mejor que la referencia de palabras clave. SVM obtuvo mayor precisión cuando utilizó unigramas como atributos (82.2%), mientras que Naive Bayes y Máxima entropía alcanzaron una precisión mayor cuando combinaron unigramas y bigramas, ambos obtuvieron 82.7%.

En Argawal *et al.,* (2011) realizaron dos tareas de clasificación: 1) clasificaron en positivo y negativo, 2) clasificaron en negativo, positivo y neutro; esto con el objetivo de ver si en alguna de las dos tareas el desempeño era mejor. Utilizaron tres modelos: modelo de unigramas, modelos de *senti-features* y modelos de árboles de Kernel (cfr. Mannig, 199: 578). El modelo de *senti-features* utiliza características de polaridad y no polaridad. Las características de polaridad son palabras de afecto

extraídas del *Dictionary of Affective Language* (DAL) (Whissel, 1989) extendidas con *WordNet* (Miller, 1995), además se utilizaron emoticones, *hashtags*, palabras con signos de exclamación asociadas a palabras de afecto. Las características no polares contienen las estadísticas de las partes de la oración que se refieren a palabras sin polaridad, *hashtags* y URLs. Además de los modelos antes mencionados, combinaron algunos: unigramas con características y características con modelos. Para cada experimento utilizaron Máquina de Vectores Soporte (SVM) y reportaron un promedio de los resultados de 5 validaciones cruzadas. Esto quiere decir que el corpus se separa en cinco partes, una se reserva para probar el clasificador y el resto se utiliza para entrenarlo. Este proceso se repite otras cuatro veces utilizando las otras cuatro partes como corpus de prueba (Manning, 1999: 586). De acuerdo con sus resultados, el modelo de árbol de kernel y el modelo basado en *senti-features* mejoran el modelo de unigrama. Para el modelo de *senti-features* las características más importantes son las que tienen polaridad y partes de la oración.

Bakliwal (2012) propone un método para determinar "la calificación popular" a partir del nivel individual de la palabra de un *tweet*. A cada *token* se le dio dos polaridades: la una positiva y otra negativa. Sumaron las polaridades positivas y negativas de cada unigrama constituyente y usaron la diferencia para encontrar toda la calificación del *tweet*. Utilizaron la frecuencia de ocurrencia de una palabra en un corpus positivo o negativo para decidir el peso de su calificación popular. Multiplicaron el factor popular (pf) con la calificación de cada unigrama *token*. También utilizaron el método de *Naive Bayes* en dos corpus distintos: el *Mejaj Dataset* (Bora, 2012) y el *Standford Dataset* (Go, 2009), el primero utiliza emoticones como etiquetas ruidosas y el segundo, palabras de sentimiento. Alcanzaron una

precisión de 87% en el *Standford Dataset* usando el *scoring method*, y 87% usando el SVM clasificador. En el *Mejaj* mejoraron un 4.77% en comparación con Bora (2012).

Las investigaciones anteriores utilizaron enfoques supervisados, es decir, utilizaron un corpus etiquetado tanto para entrenamiento como para prueba. En cambio, Montejo *et al.*, (2012) realizaron un enfoque de aprendizaje no supervisado para la detección de polaridad en *tweets*, es decir, un corpus no etiquetado. Utilizaron *SentiWordNet* para calcular la estimación final de la polaridad. Su método no mejora el SVM.

Las investigaciones sobre la clasificación de opiniones en *tweets* se han elaborado únicamente de manera automática, con enfoques en aprendizaje de máquina, muchos de ellos han considerado la polaridad del *tweet* tomando en cuenta el total de palabras positivas y negativas que tengan; sin embargo, aún en un texto tan pequeño como este se pueden encontrar dos entidades sobre las que se opinan, o dos aspectos de una entidad que tienen polaridades distintas. En ese sentido se considera que debe hacerse un análisis con mayor minuciosidad y por ello en este trabajo se realizó una clasificación manual que toma en cuenta estos aspectos.

4 ESTADO DEL ARTE

En este capítulo se detallará el estado de la cuestión sobre las preferencias lingüísticas relacionadas con el sexo. De los trabajos aquí expuestos, algunos poseen un enfoque sociolingüista, otros socio-pragmático y otros más, de perfilamiento automático y minería de opiniones. Este capítulo contiene tres subsecciones: en la primera se mencionará las publicaciones que encuentran las diferencias en la emotividad y otros que lo abordan como un problema de género discursivo; en la segunda, se hablará sobre las investigaciones que se centran en encontrar las diferencias en los medios de comunicación social, tales como redes sociales y blogs; en la tercera, se describirán estudios que utilizan las palabras que connotan sentimientos para perfilar sexo.

4.1 Tipo de discurso y subjetividad

Identificar los marcadores lingüísticos dependientes del sexo ha sido un tema bastante explotado desde mediados del siglo pasado, mas no faltaron los artículos basados en la intuición, para muestra Robin Lakoff (1975), en su libro *Language and Woman's place*, mencionaba que las mujeres construyen más *question tags* (5) y entonan enunciados de forma interrogativa donde quieren hacer una afirmación (6), es decir, utilizaban expresiones como las siguientes:

(5) The way prices are rising is horrendous, isn´t it?
(6) The dinner will be ready around six o´clock…?

Según la autora a causa de la necesidad de la mujer de abordar formas de cortesía en busca del respeto masculino. Además de las construcciones anteriores, las mujeres usan más descriptores de color (por ejemplo: *lavanda, aguamarina*), adjetivos vacíos (por ejemplo: *adorable*, *lindo*, *divino*, *dulce*) y expresiones como "*¡Oh fudge!*", "*¡Dear me!*", "*¡Oh dear!*". En cambio, el discurso masculino está caracterizado por

el mayor uso de groserías. Sin embargo, Lakoff señala que estas características no son intrínsecas de un sexo, sino que el habla insegura tiene que ver con factores sociales como el que la mujer haya estado subyugada durante muchos años y desde pequeñas se haga una distinción en la educación de los dos sexos: las niñas deben hablar como 'niñas', no deben gritar, deben ser condescendientes, no deben entrometerse en aspectos de política, etc.; mientras que a los niños sí se les permiten dichas conductas. Por otro lado, que ciertos adjetivos sean considerados como exclusivamente femeninos tiene que ver con que socialmente el hombre debe adoptar una postura de virilidad y usar adjetivos como "tierno" no está bien visto socialmente. La conclusión a la que llega la autora es que el lenguaje femenino ahoga la identidad personal de la mujer al negarle los medios con que expresarse con determinación y fomenta en ella expresiones que sugieren trivialidad en el contenido e inseguridad en el mismo (cfr. Lakoff, 1975: 23).

Muchos años después, Barbara Johnstone (1993) exploró la relación entre sexo y tipos de historias con el objetivo de hacer un estudio descriptivo de las prácticas discursivas de hombres y mujeres, asimismo, realizó un estudio teórico de la mujer y el lenguaje de una forma general. Para ello, examinó un pequeño corpus de historias personales y 'espontáneas' contadas por hombres y mujeres blancos, occidentales y de clase media. Descubrió que las historias de las mujeres tienden a ser sobre la comunidad y las de los hombres, sobre la competencia, puesto que ellos tienden a narrar historias donde actúan solos y logran el triunfo ante cualquier situación desfavorable. Además, se analizaron detalles extratemáticos y discurso indirecto. Los detalles extratemáticos son los elementos que incluyen los narradores para mover la historia dando identificación a los personajes, estos pueden ser especificaciones de lugar o tiempo, títulos de eventos, descriptores de objetos. Los resultados obtenidos

fueron los siguientes: mientras los hombres especifican más los lugares y el tiempo, las mujeres usan nombres personales más del doble de veces que los hombres. La autora sugirió que hombres y mujeres toman decisiones en su historia no solo porque tienen diferentes psicologías o porque pertenecen a diferentes subculturas, sino porque reflejan la diferencia que tienen en la toma del poder en el mundo social real como porque ellos también están creando activamente diferentes mundos en sus historias lo cual es al mismo tiempo reflejo y constitución de su psicología, su sociedad y su cultura fuera de las mismas.

Otro trabajo sobre sobre historias es el de Anna Janssen y Tamar Murachver (2004): "The role of gender in New Zeeland Literature: comparison across periods and style of writing". Ellas hallaron que las variables más usadas por las autoras femeninas incluyen complementos, adjetivos, preguntas, pronombres de tercera persona, adverbios adjetivales, metáforas, símiles, sentencias elípticas e intensificadores. Lo anterior apoya la noción de que las mujeres están más interesadas en interactuar con su audiencia, poseen un estilo más interpersonal y un tono socioemocional en comparación con los hombres. Los hombres, por otro lado, muestran mayor orientación de sí mismos, y más elementos cuantitativos y factuales en sus escritos a través de mayor uso de pronombres de primera persona, referencias geográficas y adverbios de foco (Janssen, 2004: 195).

En su estudio clasificaron el corpus en literatura de autores populares y literatura de autores consagrados y trataron de ver las diferencias de sexo a través del tiempo. Sus resultados indican que, con la incorporación de la mujer en ambientes que se consideraban masculinos, el estilo de las autoras contemporáneas ha cambiado en relación con las autoras tempranas; sin embargo, las particularidades que se encontraron en las escritoras tempranas no pueden ser atribuidas a las del resto de

las mujeres ya que ellas tenían una posición privilegiada (acceso a la educación, por ejemplo). Además, el género discursivo es un factor que influye en la elección de una estrategia lingüística, las preferencias de cada sexo estarán más marcadas en textos populares que en textos clásicos pues estos pretenden llegar a un público más amplio. Lo interesante aquí es que el sexo masculino tiene menos definido su estilo y se ve influenciado por los elementos propios del género discursivo, por lo cual resulta más difícil caracterizarlo.

Existen otros trabajos que trataron de ver las preferencias lingüísticas de sexos, no a través de períodos sino a través de géneros discursivos. Argamon *et al.* (2003) exploraron las diferencias de estilo lingüístico en la ficción y en la no ficción. Utilizaron un *Exponentiated Gradient* (EG), algoritmo que selecciona automáticamente las características más usuales para clasificar un documento. Este identificó un largo número de modificadores (especificadores) de sustantivos (*a, the, that, these*) y cuantificadores (*one, two, more, some*) como indicadores masculinos y pronombres *(I, you, he, she, her, their, myself, yourself, herself*), como femeninos. Las categorías de pronombre y especificador codifican información sobre "cosas" del mundo. Los pronombres envían el mensaje de que las entidades son conocidas por el lector y los especificadores proveen información sobre las mismas, por lo cual el escritor asume que el lector no las conoce. Los especificadores son más usados por los hombres, esto sugiere que los hombres prefieren especificar o indicar los objetos de los que hablan.

Por otro lado, cabe señalar que existe una fuerte relación entre características femeninas y masculinas con la ficción y no ficción: los pronombres aparecen con mayor frecuencia en la ficción que en la no ficción y, a la inversa, los determinantes

aparecen con mayor frecuencia en la no ficción que en la ficción. Los autores masculinos prefieren indicar o especificar las entidades de las que están escribiendo. Aunque no pudieron explorar el problema por medios automáticos, la examinación de los textos sugiere que el uso de determinantes refleja que los escritores masculinos mencionan clases de cosas en contraste con las escritoras femeninas quienes personalizan sus mensajes usando pronombres para ligar una mención de una persona u objeto con otras menciones (Argomon *et al.*, 2003:10).

El modelo de Argamon *et al.* fue utilizado por Ghafar Samar *et al.* (2010) para ver si este puede servir para identificar el sexo de hablantes no nativos de inglés. Para esto seleccionaron 32 fragmentos de diferentes diarios y cuatro artículos de 8 diarios de hablantes no nativos del inglés: dos de hombres y dos de mujeres. Las preferencias lingüísticas que se tomaron en cuenta en este estudio para caracterizar hombres son: *that, these, one, two, more, some, its, he, him*; las características para mujeres son: *I, you, she, her, their, myself, yourself* y *herself.* Las palabras de cada artículo fueron contadas y elaboraron una prueba estadística *t-test* para ver si las frecuencias de uso eran estadísticamente significativas. Aunque mostraron que los patrones tienen una frecuencia de uso diferentes entre hombres y mujeres, estas no son estadísticamente significativas, excepto *he* (mayor frecuencia en textos masculinos) y *their* (mayor frecuencia en textos femeninos). Las posibles explicaciónes de los resultados son: 1) las características propias de este género discursivo; 2) documentos insuficientes; 3) el hecho de que sean hablantes de L2 les impide expresar su sexo en toda su capacidad en los textos que producen.

Desde un enfoque pragmático, Nuria Lorenzo y Patricia Bou (2003) trató de destruir el estereotipo que considera que las mujeres son más corteses que los varones,

apoyado por Lakoff (1975). Ellas investigaron qué estrategias de cortesía son más usadas por cada sexo y realizaron una comparación entre el español peninsular y el inglés británico. Para esto propusieron seis escenarios donde las variables de poder y distancia social fueron controladas. El resultado obtenido fue que la mayoría de los hablantes orientó su interacción lingüística hacia la cortesía; sin embargo, cada sexo usó estrategias distintas para lograr su propósito y, además, el factor cultural influyó para elegir alguna estrategia.

En la UNAM, solo se encontró una tesis con un enfoque sociolingüístico variacionista titulada *Sociolingüística de los adjetivos calificativos en un corpus del español mexicano* (Rivera, 2015). La autora aborda el estudio de los adjetivos calificativos como una forma en la que los hablantes pueden emitir una opinión y compara su uso en dos períodos: 1970 y 2000. Los resultados que obtuvo para ambas épocas, le mostraron que este tipo de adjetivos son un buen indicador para distinguir este grupo social, pues las mujeres utilizan más adjetivos calificativos tanto en el eje positivo como negativo en ambos períodos. Si bien muchos de los adjetivos son utilizados por ambos sexos, son más productivos en el habla femenina, de este modo la autora refuta la idea de Lakoff (1975) sobre que algunos adjetivos sólo son utilizados por mujeres. Debido a sus resultados, Rivera concluye que "las mujeres llevan el discurso con mayor emotividad a través de los adjetivos" (2015: 131).

En algunas de esas investigaciones se insinúa que las mujeres son más subjetivas con base en los resultados obtenidos; por una lado, el mayor uso de adjetivos (Janssen y Murachver, 2004; Rivera, 2015) indica que hay mayor uso de valoración, y por tanto, mayor subjetividad dado que esta clase de palabras se utilizan para calificar o especificar una entidad; por otro, mayor uso de pronombres,

sobre todo de primera persona (Argomon, 2003; Johnstone, 1993), pueden indicar la inclusión en el discurso de un sujeto evaluador o afectado. En otra investigación (Argomon *et al.*, 2003), plantean la posibilidad de que esta subjetividad/emotividad no es propia de las mujeres, sino que tiene que ver más con las exigencias del discurso, hay mayor subjetividad en la ficción y en el habla espontánea. En consecuencia, resulta necesario plantear un estudio lingüístico sobre la subjetividad para distinguir sexo, no para ver si uno más subjetivo que otro sino para ver cuáles son las diferentes estrategias que cada uno prefiere para expresar su subjetividad. Esta tesis se enfoca en las opiniones y, aunque existe una tesis sobre las preferencias en el uso de adjetivos evaluativos para realizar una opinión (Rivera; 2015), esta no aborda el panorama completo, pues dichos adjetivos no son los únicos elementos lingüísticos que se pueden utilizar para emitir una valoración; también pueden emplearse algunas clases de verbos, adverbios, sustantivos, interjecciones, locuciones y enunciados fraseológicos. Además, la tesis mencionada tiene el propósito de ver el cambio diacrónico sobre el uso de dichos adjetivos, y para ello utiliza un corpus de entrevistas de 1970 y otro del 2000._En mi tesis, en cambio, se utilizó un corpus de mayor actualidad y con otras características (*Twitter*), por lo que se encontrarán cosas diferentes, incluso sobre el uso de dichos adjetivos (adjetivos nuevos, préstamos, etc.).

Debido a que se ha planteado el problema de que el género discursivo condiciona la preferencia de algunas construcciones lingüísticas, el siguiente apartado servirá para entender cómo se ha abordado el problema en las redes sociales y otros discursos electrónicos, pues ellos están más relacionados con el corpus que se utilizó para este trabajo.

4.2 Preferencias lingüísticas en los medios de comunicación social

Los medios de comunicación social como *blogs, Facebook* y *Twitter* han cobrado gran popularidad en las últimas décadas, dado que representan una fuente de fácil acceso para la investigación de las preferencias lingüísticas relacionadas con el sexo y, al mismo tiempo, un nuevo reto para los investigadores, ya que los nuevos géneros discursivos poseen características propias no sólo por el medio en que se encuentran.

En 2006, Herring y Paolillo investigaron si las características atribuidas al sexo en textos formales también se presentaban en textos informales que aparecen en la red (*weblogs*). En su artículo "Gender and genre variations in weblogs" describen los resultados obtenidos y concluyen que el sexo no es una variable significativa para las diferencias estilísticas sino el género discursivo. Para este trabajo los autores usaron dos tipos de weblogs: *dairy* y *filter;* y encontraron que el uso de un género discursivo en lugar de otro sí condicionaba la aparición de ciertas características lingüísticas independientemente del sexo.

Uno de los primeros trabajos que realizó perfilamiento automático de sexo en redes sociales fue Scheler *et al.* (2006). Los investigadores trataron de perfilar el sexo y la edad de los autores en un corpus de 71, 000 *blogs*. Ellos buscaron las características relacionadas con el estilo (partes del discurso seleccionadas, palabras funcionales y palabras de *blog* como 'lol' y 'hahaha') y con el contenido. La distribución de los *blogs* fue la siguiente: hasta los 17 años los *bloggers* son la mayoría mujeres (63%), pero entre *bloggers* arriba de los 17, las mujeres son casi la mitad (47%). Para ser imparciales, realizaron un subcorpus consistente en un número igual de *blogs* de hombres y mujeres en cada grupo de edad, al azar descartaron el exceso de documentos en la categoría más larga. En consecuencia, quedaron un total de 37,478 *blogs*, que comprenden 1, 405, 209 entradas y 295, 526, 889 palabras. Midieron

la frecuencia con que cada uno de los factores aparece en el corpus por sexo y edad. Encontraron que las mujeres utilizan más pronombres, palabras de asentir y negar, usan más palabras de contenido y *post length* que los hombres; en cambio, ellos usan más determinantes/artículos, preposiciones e *hyperlinks* que las mujeres. Las preposiciones y artículos que son usados con mayor frecuencia por hombres, incrementan su uso cuando los *bloggers* son de mayor edad. A la inversa, los pronombres, las palabras de negar y asentir y las palabras de *blog*, decrecen su utilización conforme las mujeres se hacen mayores. En resumen, las características que distinguen a hombres y mujeres también distinguen la edad.

Por otro lado, en las características basadas en contenido encontraron que los *blogs* masculinos incluyen muchas referencias a política y a tecnología. Casi todas las palabras incrementan o disminuyen con la edad. Las palabras que incrementan con la edad son palabras "masculinas" y las que decrecen son "femeninas". Las características tanto estilísticas como de contenido se utilizaron para clasificar de manera automática los *blogs* por edad y sexo, con la conjunción de ambas características se logró una precisión de 80.1%; sin embargo, las características estilísticas fueron de mayor ayuda que las diferencias de contenido.

En 2013, Schwartz *et al.* trataron de ver las preferencias lingüísticas en relación con tres variables: sexo, edad y personalidad. Replicaron análisis previos de vocabulario cerrado, es decir, donde existe una lista de palabras o categorías *a priori*, con análisis de vocabulario abierto, donde lo que cuenta es la frecuencia de las palabras en sí. Por un lado, se demostró que con un análisis de vocabulario abierto se pueden descubrir relaciones no encontradas previamente y se obtiene mayor precisión en los resultados que en los modelos predictivos. Las características que se

extrajeron fueron palabras, frases y tópicos de un corpus de 19 millones de estados de *Facebook* escritos por 136 mil participantes. Los resultados obtenidos fueron similares a estudios previos: las mujeres usaron más palabras de emoción (por ejemplo: *emocionado*), y pronombres de primera persona singular, y ellas mencionan más procesos psicológicos (por ejemplo: *te amo*). Los hombres usan más malas palabras, y referencias de objetos (ejemplo: *xbox*). Notaron que "*my wife*" y "*my girlfriend*" estaban fuertemente relacionados con los resultados masculinos, mientras que "*husband*" y "*boyfriend*" fueron más predictivos para las mujeres porque las mujeres eran más propensas a precederlos con adjetivos como "*sorprendente*" y adverbios, lo cual sugiere que los hombres prefieren el posesivo "*my*".

Ikeda *et al.* (2013) propusieronun método híbrido de perfilamiento automático con base en referencias de texto y referencias de la comunidad, debido a que en muchas ocasiones, algunos usuarios sí suelen hacer referencias sobre su información personal (edad, sexo y área de residencia) mientras que otros prefieren ocultarla. El método con base en referencias de la comunidad utilizó los seguidores y seguidos de cada usuario, ya que las personas tienden a compartir información con sus compañeros de trabajo, personas de la misma edad o con quienes comparten los mismos pasatiempos. En la fase de entrenamiento del método de referencias de texto, primero fue analizado el historial de *tweet* de usuarios conocidos, después fueron extraídos los términos característicos y finalmente, fueron entrenados los SVMs[3] con las características de los términos usados. Utilizaron un AIC[4] para ver la relevancia

[3] Máquina de vectores soporte, para mayor información cfr. Manning, 1999: 589, 539, 43.

[4] *Akaike's Information Criteria* es una pauta para la evaluación de modelos estadísticos. Se utilizó para comparar el modelo dependiente (la aparición del término t está relacionado con la demografía) y el modelo independiente (la aparición del término t es independiente de la demografía). Véase Ikeda, 2013: 38.

de cada término y así poder entrenar los SVM. La aparición de un término podía tener dos relevancias una positiva y otra negativa, la relevancia positiva significa que el término aparece frecuentemente en un segmento demográfico etiquetado, el cual correspondía con un sexo, un grupo etario, un estado civil, una ocupación o un *hobbie.* La relevancia negativa significa que el término aparece frecuentemente en *tweets* de usuarios no etiquetados en ese segmento demográfico. Los vectores de características fueron extraídos de los *tweets* de los usuarios. Los vectores de características de un usuario fueron construidos basados en qué términos de la lista de términos del segmento demográfico aparecen en el historial de *tweets* del usuario. En la fase de estimación, fueron extraídos los vectores de características de usuarios desconocidos y fue estimada la probabilidad de que un usuario pertenezca a un segmento demográfico por su historial de *tweets*.

El método de base de referencias de la comunidad analizó la relación entre los seguidores y los seguidos en usuarios desconocidos con el fin de extraer comunidades. Una comunidad se inicia con un usuario etiquetado, a este se añaden sus vecinos hasta que el número de la comunidad es igual a un límite determinado, en el experimento probaron con 100, 300, y 500 debido a que cuando la comunidad es muy pequeña no refleja sus características, pero cuando es muy grande se incluyen usuarios no relacionados. La demografía de las comunidades extraídas son estimadas con el método base texto.

Con una API, fueron recolectados *tweets* con caracteres japoneses. Buscaron los términos que dijeran la demografía de los usuarios. En total fueron recolectados 100,000 perfiles de usuarios que fijaron su edad, de los cuales fueron seleccionados 600 para cada segmento demográfico, a su vez cada segmento fue dividido en dos grupos: uno se utilizó para entrenamiento y otro como prueba. Se hicieron 5

repeticiones de dos grupos de validación cruzada. El método híbrido funcionó adecuadamente para edad, ocupación y *hobbies* y área. Sin embargo, este fue inválido para género ya que ambos sexos tienen muchas redes con sus compañeros de trabajo y amigos, y por lo tanto, las comunidades que se forman a partir del método con base comunidad siempre estarán formadas tanto por hombres como por mujeres.

En este apartado, dos de los trabajos vuelven a encontrar más elementos asociados con la subjetividad para distinguir a las mujeres: Scheler (2006) encuentra mayor uso de pronombres y Schartz (2013), además de pronombres de primera persona, encuentra palabras de emoción y procesos psicológicos, es decir, palabras que expresan estados privados, y adjetivos.

4.3 Perfilamiento automático y sentimientos

Los trabajos revisados en el apartado anterior son una pequeña muestra del estado del arte sobre perfilamiento automático con diferentes características y extracción de características para diferenciar sexo en diversos medios sociales. En este espacio, se revisarán los trabajos que utilizaron como rasgos diferenciadores las palabras de sentimientos. Hasta el momento, se encontraron tres artículos: Patra *et al.* (2013), Weren *et al.* (2014) y Pimas *et al.* (2016).

Patra *et al.* (2013) utilizaron el corpus del PAN 2013 para perfilar sexo y edad. Ellos llevaron a cabo la tarea para el inglés y, además de las palabras de sentimiento (positivas y negativas), consideraron los pronombres, los temas de las palabras, las palabras funcionales, los extranjerismos, los emoticones, signos de puntuación y el promedio de palabras por cláusula. Primero, realizaron una limpieza del corpus, donde eliminaron palabras ininteligibles, después, calcularon los emoticones y las palabras funcionales. Luego, delimitaron la oración y calcularon la

cantidad de palabras por enunciado. Más adelante, etiquetaron sus categorías gramaticales. Al final, calcularon las frecuencias de las clases de palabras, las palabras negativas y positivas, los signos de puntuación y los pronombres. Para el entrenamiento, utilizaron un árbol de decisiones y obtuvieron una precisión máxima de clasificación de 56.83% para sexo y 28.95% para edad.

Weren *et al.* (2014) también realizaron perfilamiento para sexo y edad. En su caso, las palabras de sentimientos no se limitaban a expresar sentimientos positivos o negativos, sino que también analizaron qué tipo de emoción estaba asociada (sorpresa, alegría, miedo, tristeza, molestia, disgusto, anticipación, confianza) puesto que se utilizó el *NRC Emotion Lexicon* (Mohammad y Turney, 2013). Además de estas palabras, utilizaron como características la longitud de oración (cantidad de palabras por oración), la longitud de palabra (cantidad de caracteres por palabra), la longitud de párrafo (cantidad de oraciones por párrafo), legibilidad del texto (la dificultad de comprensión de un texto), qué tan bien escrito estaba el texto, es decir, repetición de vocales, puntuación y palabras mal escritas y la diversidad de las palabras (número de palabras distintas/ número de palabras en el texto). En primer lugar, realizaron un pre-procesamiento del texto: removieron los caracteres repetidos; posteriormente, utilizaron diversos clasificadores como SVM y árboles de decisiones.

Pimas *et al.* (2016) realizaron una tarea de perfilamiento automático para edad y sexo a través de varios textos en inglés (*tweets*, *blogs,* posts en los medios sociales), esto es, utilizaron un tipo de texto para entrenar y probaron con otro. Su trabajo utilizó un clasificador de *bosque aleatorio* y como características: concreción, sentimientos e información sintáctica. Por concreción se refieren a palabras que son fáciles de recordar a diferencia de los conceptos abstractos. Todas las palabras

poseían tres calificaciones: el significado de concreción, la desviación estándar de la concreción y el porcentaje de conocimiento.

Para la primera calificación, los anotadores evaluaron la concreción de las palabras en un *tweet* de 5 a 1. La desviación estándar codifica qué tan fuerte es el acuerdo de concreción entre los anotadores. Para las palabras en las que todos los anotadores están de acuerdo, la calificación es baja. El porcentaje de conocimiento representa el porcentaje de los evaluadores que indicaron conocer la palabra. Para encontrar las características a nivel de *tweet,* todas las calificaciones basadas en las palabras fueron agregadas. Por lo tanto, el mínimo, el máximo y el significado aritmético fueron computados para cada uno de los tres tipos de calificaciones.

Para las características de sentimiento utilizaron la librería *SentiWordNet*, con la cual extrajeron la calificación de cada palabra. Obtuvieron para cada *tweet* la máxima polaridad, la mínima polaridad, el promedio de la polaridad y la desviación estándar de la polaridad de todos los términos con polaridad.

Como rasgos sintácticos utilizaron la extensión de la palabra (esperaban más acrónimos y abreviaciones) y el uso de *hashtags*. Dichos elementos fueron considerados por las propiedades de *Twitter*.

La precisión en la clasificación fue inconsistente, en un primer corpus obtuvieron 0.5769 para sexo, adecuado para el estado del arte, y en el segundo, 0.0201, un resultado bastante bajo.

Finalmente, se consideró importante incluir en esta sección, un trabajo que no trata de realizar perfilamiento automático sino que muestra qué emoción predomina en los textos de hombres y mujeres. Mohammad and Yang (2011) investigaron hacia qué emoción orientan sus palabras hombres y mujeres de cartas de amor, correos de

odio y notas suicidas. Para ello, elaboraron un lexicón de sentimientos a partir del *Rogert Thesaurus*. Sólo se anotaron las palabras que ocurrían más de 120,000 veces en el *Google n-gram corpus*. Los cinco anotadores tenían que responder si la palabra estaba asociada con un sentimiento positivo o negativo: enojo, miedo, alegría, tristeza, disgusto, sorpresa, confianza y anticipación. Para las palabras con un sólo sentido tomaron en cuenta la emoción asociada a la mayoría de votos, mientras que para las palabras que tuvieran diferentes sentidos tomaron todos los sentimientos asociados. Dando como resultado el *NRC Emotion Lexicon 0.92* (Mohammad y Turney, 2013) con 14, 200 tipos de palabras. Para cada texto del corpus, su sistema determinaba qué palabras existían en el diccionario de emociones y calculaba el promedio de palabras relacionadas con una emoción en relación con el total de palabras de emoción de un texto. Este enfoque sirvió para determinar qué emoción está más presente en un texto. Descubrieron que las mujeres utilizan más palabras que se encuentran en el eje de la alegría y la tristeza, mientras que los hombres se ubican en el eje de la verdad y el miedo. Su trabajo apoya la idea que los hombres y mujeres usan las palabras de emoción en diferentes grados.

Como pudo notarse, en los tres trabajos que utilizan palabras de sentimientos para perfilamiento automático, no se encontró que estas fueran de gran ayuda para clasificar sexo (Patra *et al.*, 2013; Weren *et al.*, 2014; Pimas *et al.*, 2016), y sin embargo, en el trabajo de Mohammad and Yang (2011) se pudo notar que las emociones predominantes en un texto masculino y femenino son diferentes. Probablemente, debido a la falta de conocimiento lingüístico por parte de los computólogos para discriminar qué clases de palabras de sentimientos utilizar o en qué función, en lugar de utilizarlas todas.

La mayoría de los trabajos encontrados se ha realizado para el idioma inglés (Lakoff, 1975; Johnstone, 1993; Janssen y Murachver, 2004; Argomon *et al.*, 2003; Scheler *et al.*, 2006; etc.), con excepción de Rivera (2015) que se hizo para el español y Lorenzo y Bou (2002). Lo anterior indica que hasta el momento ha habido mayor interés en el inglés para perfilar sexos.

Existen varios trabajos que realizan perfilamiento de autor en las redes sociales y *blogs* (Herring y Paolillo, 2006; Scheler *et al.*, 2006; Schwartz *et al.*, 2013; Ikeda 2013) pero pocos que han utilizado las palabras afectivas para determinar género (Patra *et al.*, 2013; Weren *et al.* 2014; Pimas *et al.*, 2016), y quienes las han usado no han tenido buenos resultados con estas características, pese a que se atribuye mayor subjetividad al discurso femenino por el uso de pronombres (Johnstone, 1993;), adjetivos (Lakoff, 1975; Janssen y Murachver, 2014; Rivera, 2015; Schwartz *et al.*, 2013) y procesos psicológicos (Schwartz *et al.*, 2013).

Por otro lado, el discurso masculino ha sido caracterizado por el uso de cuantificadores, especificadores de lugar y tiempo (Argomon *et al.*, 2003; Janssen y Murachver, 2004; Johnstone, 1993; Scheler *et al.*, 2006) y groserías (Lakoff, 1975; Schwartz *et al.*, 2013).

5 PALABRAS Y FRASES DE OPINIÓN

En el capítulo tres ya se mencionó que algunas palabras son un fuerte indicador de opiniones y por ello se crean diccionarios de sentimientos que permiten identificarlas de manera automática. En este capítulo, se discutirá sobre las palabras y expresiones que contienen sentimientos, es decir, se darán las definiciones y funciones de las clases de palabras (sustantivos, adjetivos, adverbios, verbos, interjecciones) y frases (locuciones y enunciados fraseológicos), así como algunas de las clasificaciones semánticas que permitieron que muchas de ellas se considerarán como palabras de opinión.

5.1 Unidades de opinión o subjetivemas

Kerbrat Orecchioni (1986) propuso el término subjetivemas para designar a las unidades léxicas con rasgos afectivos y axiológicos, con el último atributo se refería a los vocablos que poseían una valoración del tipo positivo- negativo.

El valor axiológico de un término puede ser más o menos estable o inestable. Existen términos que poseen una connotación positiva o negativa dentro del sistema lingüístico; pero también, los valores axiológicos dependen del idiolecto de cada persona y de los valores culturales atribuidos a un término o tema, por ejemplo, en la juventud y la vejez, puesto que en algunas culturas la vejez significa sabiduría y la juventud, falta de experiencia, y en otras donde se trata de eliminar cualquier rasgo se senectud porque esta es desagradable.

Los subjetivemas entonces pueden ser adjetivos, sustantivos, verbos, adverbios e interjecciones que posean un rasgo axiológico o sentimental dentro de su significado. En los subsecuentes apartados se explicará cada una de esas unidades.

5.1.1 Sustantivos

Los sustantivos son una clase de palabra que se caracteriza por admitir género y número, y formar grupos nominales que cumplen diversas funciones sintácticas como el sujeto, complemento directo, término de preposición, aposición, etc. (cfr. RAE, 2010: § 1.2.1.1). Desde un criterio semántico, pueden definirse como "nombres con que se designan los seres que son objeto de nuestros juicios." (Seco, 1967: 152), o dicho ampliamente,

> denotan entidades, es decir, realidades percibidas por el hablante como diferenciadas. Estas entidades pueden ser objeto del universo físico (*mano*, *corazón*, *olor*), acciones (*lectura*, *movimiento*, *pelea*), acontecimientos (*muerte*, *caída*, *destrucción*), situaciones (*pobreza*, *paz*, *vejez*), realidades culturales (*música*, *gastronomía*, *filatelia*), realidades conceptuales (*idea*, *ecuación*, *teoría*), realidades del mundo psíquico (*emoción*, *manía*, *talento*), relaciones (*amistad*, *obediencia*, *cuñado*), cualidades (*belleza*, *velocidad*, *simpatía*), etc. Prácticamente cualquier aspecto de la experiencia humana puede ser expresado por medio de un sustantivo. (Fages, 2005: 74-75)

Existen muchas formas de clasificar los sustantivos, entre ellas: los sustantivos comunes (los cuales refieren a grupos de entidades como son: *elegancia*, *niña*) vs propios (nombran entidades únicas y particularizadas: *México*, *Juan*); a su vez, dentro de los comunes se encuentran los concretos (designan entidades que pueden ser apreciadas por los sentidos, por ejemplo: *hoja*, *humo*) vs abstractos (refieren a entidades no perceptibles: *verdad*, *amor*, *felicidad*).

De acuerdo con Kerbrat (1986), existen sustantivos que connotan un juicio positivo o negativo, estos se denominan sustantivos afectivos y evaluativos o axiológicos, de los cuales, algunos pueden ser derivados de verbos o adjetivos, tales como *amor*, *belleza*, etc. Los sustantivos que proceden de verbos conservan algunas de las características de ellos como el número de argumentos. Los sustantivos

axiológicos son aquellos que se encuentran en el terreno de lo peyorativo (desvalorizador)/elogioso (laudatorios, valorizadores) y proporcionan tanto una descripción del denotado como un juicio evaluativo que hace el sujeto de la enunciación sobre él. Tal es el caso de:

(7) Alguien es una *mierda*, un *canalla*, etc.

Dixon (1991) propone una clasificación semántica del sustantivo en 5 clases: referencias concretas, referencias abstractas, actividades, actos de habla y estados (y propiedades). Este último cubre los estados mentales (*placer*, *honor*, *diversión*, *alegría*, *habilidad*, *sagacidad*) y los corporales. Algunos son básicos como (*enojo* y *hambre*) pero muchos son derivados de adjetivos (*jealousy*: *celos*) y otros de verbos (*delight*: *deleite*).

5.1.2 Adjetivos

Los adjetivos son la clase de palabra que se adjunta al sustantivo para calificarlo o determinarlo. Formalmente, se caracterizan por tener género y número dependiente del sustantivo al que modifica (*alto/alta/altos/altas*), aunque hay adjetivos que sólo tienen flexión de número pero no de género (*posible/posibles*, *audaz/audaces*) y otros invariables (*gratis, antiarrugas*) (cfr. Fages, 2005). Dentro de esta categoría se encuentran los calificativos y los relacionales. Los primeros, como el nombre sugiere, designan cualidades en sentido estricto y los segundos, indican propiedades que la entidad objeto de modificación adjetiva posee por su relación con algo externo a ella (cfr. Bosque y Demonte, 1999).

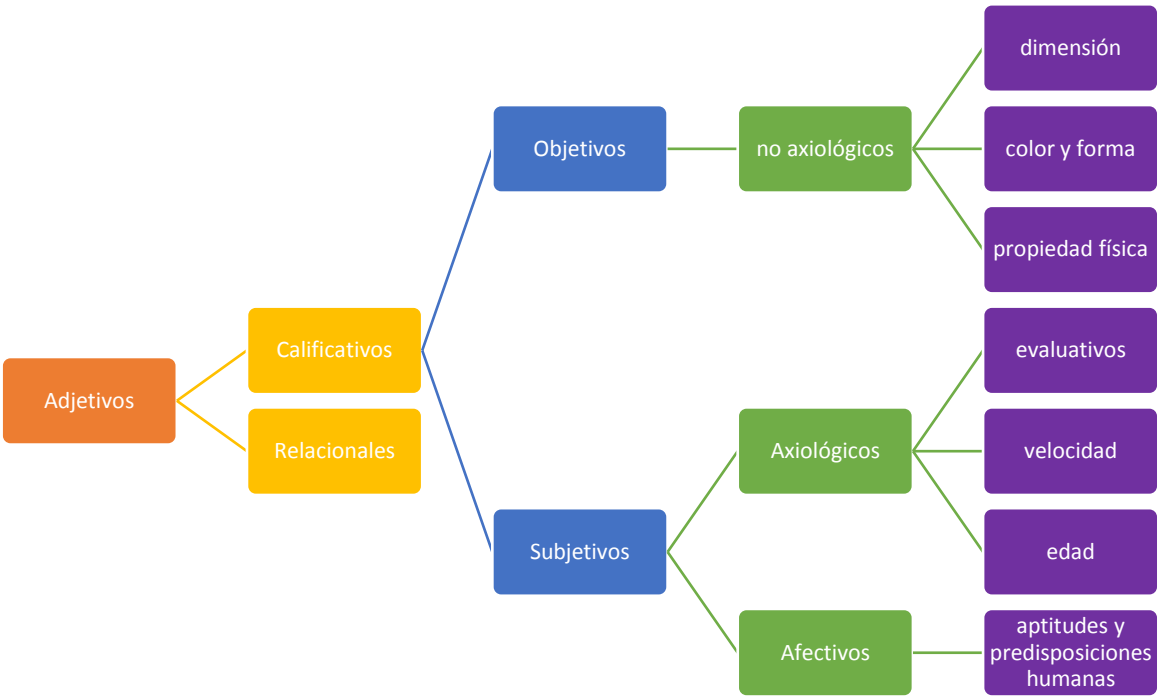

Figura 1. Clasificación semántica de los adjetivos

Los adjetivos calificativos son comúnmente utilizados para emitir una valoración sobre una entidad, puesto que suelen designar las cualidades del sustantivo al que acompañan; sin embargo, no todos los calificativos son subjetivos. En la figura 1 se muestra una síntesis de las tres clasificaciones semánticas de los adjetivos de las que se hablará a continuación. Dentro de los adjetivos calificativos, Fages (2005: 122) señala dos subtipos (recuadros azules):

a. **Objetivos:** presentan la propiedad del sustantivo como independiente de cualquier observador: forma, color, peso, tamaño, situación, etc.: una mesa blanca, una caja abierta, un melón maduro, un camino oculto.
b. **Subjetivos:** presentan la propiedad del sustantivo ligada a la opinión del hablante: una mesa preciosa, un presentador simpático, un premio deseado, una noche espantosa, etc.

La definición parece bastante clara, pero en la realidad, un adjetivo que se refiera al color o tamaño de una entidad puede ser muy subjetivo, porque lo que para

una persona es pequeño, para otra puede ser muy grande. Sin embargo, aunque estas entidades puedan ser juzgadas como subjetivas desde esa perspectiva, para este trabajo no fueron consideradas debido a que no proporcionan una valoración positiva o negativa sobre la entidad.

A los adjetivos subjetivos, Kerbrat (1986: 11-120) los divide en tres categorías (rectángulo verdes figura 1): los evaluativos no axiológicos, los axiológicos y los afectivos. Los primeros (*grande*, *lejano*, *caliente*, *abundante*) son aquellos que sin enunciar un juicio de valor, ni un compromiso afectivo del emisor, implican una evaluación cualitativa o cuantitativa del objeto denotado por el sustantivo al que determinan. El autor los considera subjetivos debido a que una entidad se mide con respecto a una unidad que cada sujeto piensa y son equivalentes a los que Fages (2005) llama objetivos. Los adjetivos axiológicos (*bueno*, *lindo*, *correcto*) aplican un juicio de valor, positivo o negativo al objeto que denota el sustantivo. Los afectivos (*desgarrador*, *alegre*, *patético*) dicen una propiedad del objeto al que determinan y una reacción emocional del hablante con respecto a ese objeto.

Existen algunas similitudes entre los valores afectivo y axiológico, entre los mecanismos psicológicos de participación emocional y de (des)valorización. Las dos clases no coinciden, pero se intersectan en algunos términos (admirable, despreciable, excitante, irritante) y estos son considerados como axiológico-afectivos (Kerbrat, 1986).

De la clasificación de Kerbrat Orecchioni sólo se consideraron los adjetivos axiológicos y los afectivos, pues sólo ellos presentaban una valoración, los no axiológicos, no.

Existe una clasificación semántica nocional de los adjetivos en inglés propuesta por Dixon (1977) y adaptada al español por Bosque y Demonte (1999) (rectángulos morados figura 1):

A. **Adjetivos de dimensión**. Considera las tres dimensiones espaciales de los objetos físicos (largo, ancho y volumen o profundidad) *largo*, *corto*, *alto*, *bajo*, *ancho*, *estrecho*, *angosto*.

B. **Adjetivos de velocidad**. *Rápido*, *lento*, *veloz*.

C. **Adjetivos de propiedad física**. Hacen referencia a propiedades de los objetos perceptibles mediante los sentidos. Entre estas propiedades encontramos forma, peso, consistencia, sabor, tacto, olor, temperatura, sonoridad.

D. **Adjetivos de color y forma**. Compuestos sintagmáticos como *rojo granate* y *verde botella*.

E. **Adjetivos de edad**: *viejo*, *nuevo*, *joven*, *arcaico*.

F. **Adjetivos de evaluación o evaluativos**. Se incluyen los aspectos de la realidad, humana y no humana, que los seres humanos consideran susceptibles de valoración. No es una clase más entre los adjetivos calificativos sino más bien una hiperclase que cruza a casi todas las anteriores.

G. **Adjetivos de aptitudes y predisposiciones humanas**. Aptitudes emocionales (*sensible*, *amable*, *cordial*, *cariñoso*), intelectuales (*inteligente*, *capaz*, *sabio*), y pasiones y disposiciones primordiales (*nervioso*, *agresivo*).

Ya que se expusieron algunas de las clasificaciones semánticas de los adjetivos, es necesario describir sus funciones y la capacidad que tienen para ser graduados, pues esta puede intensificar o proporcionar una opinión contraria.

5.1.2.1 Graduación de los adjetivos

Una de las características de los adjetivos calificativos es que permiten la graduación y con ello "la matización de nuestras opiniones" (Romero, 1989: 62). Esta puede realizarse anteponiendo un adverbio intensificador como *muy*, *tan*, *extremadamente*, etc.

Los adjetivos calificativos tienen la capacidad para conformar el comparativo y el superlativo. Pueden considerarse dos tipos de comparativos: de desigualdad o de igualdad. El comparativo de igualdad puede ser de superioridad o de inferioridad y se conforma de la siguiente manera (cfr. Romero, 1989: 67-68):

más + adjetivo base + que (superioridad/desigualdad)
A es más tierno que B
menos + adjetivo base + que (inferioridad/desigualdad)
A es menos divertido que B

El comparativo de igualdad se conforma así:

tan+ adjetivo base + como (igualdad)
A es tan amargo como B

El superlativo o comparativo de excelencia, el cual indica superioridad o inferioridad sobre todo un grupo de elementos, puede formarse de dos maneras: morfológica o léxico sintáctica. Con la primera se utiliza la sufijación (-ísimo, -érrimo), la prefijación (archi-, re-, super-, hiper-). La segunda se realiza con la siguiente estructura (Romero, 1989:64-67):

artículo + más/menos + adjetivo base

A es la más grande

5.1.2.2 Funciones sintácticas de los adjetivos

La función sintáctica del adjetivo depende de la posición que tenga con respecto al sustantivo (cfr. Romero, 1989). Las principales funciones sintácticas son la atributiva y la predicativa.

La función atributiva sucede cuando el adjetivo "incide directamente sobre el sustantivo" (Romero,1989: 89), extendiendo el sintagma nominal. Por ejemplo:

(8) La *hermosa* casa *azul* donde solía vivir.

En esta función, el adjetivo puede colocarse antes o después del sustantivo. Los adjetivos calificativos de caracterización física y moral "son frecuente expresión de valoraciones personales y emotivas y, por lo tanto, los que más a menudo preceden al sustantivo" (Rafael Lapesa, 1989: 92). Es por ello que a la anteposición suele ligarse con "matices especiales ligados a la afectividad" (Romero, 1989: 97).

Por otro lado, los adjetivos relacionales, aquellos que no expresan una cualidad sino un elemento externo con el que se relaciona una entidad, se posponen. La posposición está ligada con la objetividad; sin embargo, cuando se trata de adjetivos calificativos evaluativos "la valoración se mezcla con lo descriptivo" (Romero, 1989: 92).

En cuanto a la función predicativa, podemos decir que en ese caso "el adjetivo está unido al sustantivo mediante una cópula verbal explícita o implícita" (Romero, 1989: 97), es decir, se encuentra en el sintagma verbal (Fages, 2005). Por ejemplo:

(9) La casa azul **era** *hermosa.*

El adjetivo también se puede nominalizar o sustantivar anteponiendo los artículos *el*, *la*, *los*, *las*, *lo* (Romero Gualda, 1989). Cuando se utiliza el artículo *lo*, el adjetivo conserva la característica de gradación y puede llevar un intensificador. Ej.:

(10) Hay que resolver **lo más difícil.**

El proceso de sustantivación inicia con la supresión de un sustantivo de significado general como persona, cosa, sujeto, individuo, hombre, mujer, y, mientras la sustantivación esté en proceso siempre se puede devolver su valor original al sustantivo si se inserta un sustantivo al sintagma.

(11) Vino una [mujer] joven.

Existe otra función del adjetivo que no es sintáctica sino discursiva: el vocativo. Estos aparecen en el texto escrito como palabras separadas del resto del enunciado por comas y con signos de admiración; en el lenguaje hablado requieren mayor intensidad y entonación especial. Pertenecen a la función apelativa del lenguaje. De acuerdo con Gili Gaya (1993[1943]: 214): "El adjetivo no es complemento de ninguno de los componentes de la oración, ni guarda con ellos relación gramatical alguna. Por esto va sin preposición [...] son como las interjecciones". Ejemplo:

(12) Hola, **linda**, ¿cómo estás?

5.1.3 Verbos

Los verbos denotan acciones (*saltar*, *correr*), transformaciones (*crecer*, *perder*) o estados (vivir, saber) que se producen en un periodo. Sus características son (Fages, 2005: 152-153):

- Acepta morfemas flexivos de persona y número, los cuales marcan una persona gramatical distinta *llor-ab-a*. Existen formas finitas (verbos conjugados) y no finitas (infinitivo, participio y gerundio).
- Acepta morfemas flexivos de tiempo, los cuales sitúan los procesos con respecto al momento de la enunciación. Algunos tiempos verbales son: presente, pretérito imperfecto, futuro perfecto, pretérito indefinido, etc.
- Acepta morfemas flexivos de modo para expresar la forma en que el hablante concibe el proceso designado por el verbo. Este proceso puede ser contemplado por el hablante de tres formas básicas: a) como un proceso real, indicativo, b) como un proceso hipotético, subjuntivo, c) como un proceso en el que el hablante quiere que se haga real, modo imperativo.
- Acepta morfemas flexivos de aspecto que denotan la forma en que se produce internamente el proceso nombrado por el verbo. Tiene que ver con cuestiones como el inicio y el final del proceso, el hecho de que éste sea continuado o esté formado por momentos sucesivos, la forma en que el sujeto participa en la realización, etc. Por medio de la morfología verbal, el único aspecto claro que puede expresarse en español es el que indica el carácter perfectivo o imperfectivo del proceso, es decir, la diferencia entre el proceso visto sin considerar su momento final y el proceso visto como finalizado.

Los verbos pueden ser copulativos o predicativos. Los verbos copulativos (*ser, estar, parecer*) actúan como fragmento de unión entre el sujeto de la oración y el atributo, mientras que los verbos predicativos poseen un significado léxico y proporcionan información sobre el sujeto.

Kerbrart Orecchioni (1986) realiza una distinción de los verbos subjetivos en verbos ocasionalmente subjetivos e intrínsecamente subjetivos. Para esta clasificación considera tres aspectos: 1) quien hace el juicio evaluativo, es decir, si es el locutor o el agente del proceso; 2) cuál es la entidad que se evalúa el proceso mismo o un objeto del proceso: una cosa, un individuo o un hecho, y 3), cuál es la naturaleza del juicio,

esto es, si está en el dominio de lo axiológico (bueno o malo) o de la modalización (verdadero/falso/incierto).

Los verbos ocasionalmente subjetivos implican una evaluación del objeto del proceso por parte del agente del proceso en términos de bueno/malo o de verdadero/falso. Llamados "verbos de modalidad", "verbos modificantes" o "verbos evaluativos de actitud proposicional" expresan la actitud de un sujeto frente a un objeto. Sin embargo, no poseen juicio de valor si no están conjugados en primera persona.

Cuando la evaluación es del tipo bueno/malo se encuentran verbos de sentimientos y verbos de 'decir'. Los verbos de sentimiento expresan una disposición favorable o desfavorable del agente del proceso frente a su objeto y, al mismo tiempo, una evaluación positiva o negativa de este objeto. Entre este tipo de verbos están *gustar*, *querer*, *apreciar*, *desear*, *ansiar*, *amar*, *odiar*, *detestar*, *subestimar*, *temer*, *lamentar*, *menospreciar* y *aborrecer*. En los verbos de 'decir', el estado afectivo de 'x' se explícita en un comportamiento verbal. Entre ellos están: *lamentar*, *deplorar*, *quejarse de*, *alabar*, *censurar* y *denostar*.

Cuando la evaluación es del tipo verdadero/falso/incierto se utilizan los verbos que denotan la manera como un agente aprehende una realidad perceptiva o intelectual más o menos segura o más o menos discutible. Si la aprehensión es perceptiva, los verbos funcionan como índice de subjetividad e indican que la impresión perceptiva es específica del individuo que la recibe (*le parecía*, *tenía la impresión*). Si la aprehensión es intelectual, se emplean los verbos de opinión y sirven al locutor para informar al destinatario acerca de las opiniones de un tercero, además indica cuál es el grado de certeza con que ese tercero se adhiere a su opinión (pensar, creer, opinar, saber).

Los verbos intrínsecamente subjetivos implican una evaluación cuya fuente siempre es el sujeto de la enunciación. También se pueden subclasificar dependiendo del tipo de naturaleza de la evaluación: si es bueno o malo, si es verdadero, falso o incierto.

Cuando la evaluación es del tipo bueno/malo, el verbo implica una evaluación hecha por el locutor sobre el proceso denotado de naturaleza axiológica en términos desvalorizadores. Se utilizan verbos como *ulular*, *graznar*, *vociferar*, *aullar*, *rebuznar*, *bablear*, *heder*, *apestar*, *perpetrar*, *reincidir*, *infligir*, *resentirse*, *fracasar*, *triunfar*, *dedicarse*, *revolcarse en*, *degenerar*, *retroceder*, *merecer*, *soportar*.

Cuando la evaluación es del tipo verdadero/falso/ incierto, se utilizan verbos de 'decir' y opinar que contienen la actitud intelectual de 'z' frente a 'p'. Tal es el caso de *decir*, *afirmar*, *declarar*, *sostener*, *pretender*, *reconocer*, *confesar*, *admitir*, *pretextar*, *contradecirse*, *jactarse*, *saber*, *ignorar*, *sospechar*, *imaginar*, *piensa*, *creer*.

Para Dixon (1991), dentro de los verbos existe una clase para expresar molestia (*Depress, scare, terrify, worry, sadden, madden, concern, infuriate, annoy*) y otros para expresar gusto (*like, love, prefer, enjoy, favour, etc*). Estos tipos de verbos tienen un experimentador (generalmente humano) quien tiene una experiencia interna provocada por un estímulo.

A este tipo de verbos también se les conoce como verbos psicológicos pues "hacen referencia a estados de ánimo y/o emociones o sentimientos que producen en el hablante los hechos, acciones o situaciones que enuncia (temor, placer, duda, gratitud, sorpresa, molestia, desagrado, sorpresa, extrañeza, indignación)" (Sastre Ruano, 1997:63 citado por Sedado, 2006: 898), son verbos de "apreciación y juicio de valor (con los que el hablante manifiesta su reacción ante un hecho)" (Sedano, 2006: 898) y según Belleti y Rizzi (1988) poseen un experimentador, el individuo que

experimenta el estado privado y un tema, el contenido o el objeto de un estado mental (291).

5.1.4 Adverbios

Los adverbios son una clase de palabra que no pueden tener morfemas flexivos, con excepción del de grado, en algunos casos, y se pueden clasificar de la siguiente forma de acuerdo con sus características formales:

1. **Léxicos.** Palabras invariables con ausencia de flexión de género y número. Originalmente, ninguna de ellas pertenece a una categoría distinta de la adverbial. (abajo, adelante, adrede, ahora, alrededor, anoche).
2. **Derivados.** Formados por la adjunción del sufijo mente (sabiamente)
3. **Adjetivales.** Provienen de adverbios que perdieron su capacidad flexiva (alto, bajo, rápido, lento) (cfr. Fages, 2005)

Este tipo de palabras puede funcionar como:

1. Función de intensificador de un SA o de un SAdv: El cuadro estaba ligeramente inclinado
2. Función de complemento subcategorizado de V: El chico se comportó correctamente
3. Función de modificador verbal: Juan apenas sale de casa
4. Complemento circunstancial de cualquier tipo: El coche arrancó bruscamente
5. Función de nexo o conector entre cláusulas (adverbios conjuntivos): la llamé mientras esperaba el tren (cfr. Fages, 2005: 235)

Esta categoría suele fungir como intensificador de los sintagmas adjetivales, como se vio en el apartado de los adjetivos; sin embargo, también pueden ofrecer ejemplos de unidades subjetivas afectivas y axiológicas examinadas por Kerbrat Orecchioni (1986), sobre todo, los adverbios formados con el sufijo -mente y aquellos

que provienen de adjetivos, tal es el caso de *terriblemente*, *desgraciadamente*, *difícilmente*.

5.1.5 Interjecciones

Las interjecciones "comunican sentimientos e impresiones, [...] ponen de manifiesto diversas reacciones afectivas o se induce a la acción" (RAE, 2010: § 32.1.1a). Con ellas se puede "manifestar sorpresa, asentimiento, o rechazo." (RAE, 2010: 32.1.2)

Existen dos tipos de interjecciones:

a. **Apelativas o directivas**: orientadas hacia el oyente con intención de moverlo a la acción o provocar una reacción emocional en él. Ejemplos: *¡cuidado!*, *¡ánimo!*, *¡ojo!*
b. **Expresivas o sintomáticas**: se orientan hacia el hablante en el sentido que manifiestan sus sensaciones, sentimientos y otros movimientos de ánimo. Manifiestan emociones o reacciones del que habla. Por ejemplo: *¡lástima!*, *¡maldición!*, *¡caramba!* (RAE, 2010).

Entre la interjecciones expresivas o sintomáticas se encuentran *¡Ay!*, *¡bah!*, *¡bravo!*, *¡guau!*, *¡Aj!*, *¡puaj!* Algunas de ellas manifiestan dolor; otras, indiferencia; otras más, repulsión, y otras, admiración.

5.2 Unidades fraseológicas

Además de las clases léxicas previamente expuestas, existen palabras que se agrupan y que en conjunto pueden proporcionar una valoración sobre una entidad, se les conoce como *unidades fraseológicas* (UF). Corpas (1996) utiliza el término para nombrar al conjunto de "unidades léxicas formadas por más de dos palabras gráficas en su límite inferior, cuyo límite superior se sitúa en el nivel de la oración compuesta" (1996: 20). Dichas expresiones se caracterizan por:

1. **Frecuencia de coaparición y de uso.** La frecuencia de coaparición se refiere a la cantidad de veces que los elementos constituyentes de las unidades fraseológicas aparecen de manera simultánea en una muestra de lengua en comparación con las veces que esos elementos aparecen separadamente. La frecuencia de uso se refiere a la alta frecuencia de aparición en una muestra de la lengua.
2. **Institucionalización**. Es el reconocimiento y aceptación de las UFs por la comunidad de hablantes como consecuencia de la frecuencia de uso.
3. **Estabilidad:**

    a. Fijación

    - **Interna:** restricción de la elección de los componentes, imposibilidad de reordenamiento y de contenido.
    - **Externa:** el uso de las UFs en contextos preestablecidos (despedidas, saludos) o en determinadas posiciones en el ordenamiento de textos (despedidas en las cartas, por ejemplo).

    b. **La especialización semántica** se refiere a la relación directa y unívoca de las UFs con su significado por parte de los hablantes.

4. **Idiomaticidad.** Es una propiedad semántica que está presente sólo en algunas unidades fraseológicas, con la cual el significado global de una expresión no puede entenderse por la suma de los significados de sus elementos constituyentes.

5. Variación

- **Variantes:** UFs similares en estructura y componentes que no poseen diferencias de significado. Por ejemplo: *Todo queda en casa /todo queda en familia*.

- **Variaciones** por derivación (*ser un culo/culillo de mal asiento*) o transformación (*Metedura de pata* por *Meter la pata.*)
- **Modificaciones** ocasionales y creativas de la UFs se pueden presentar en el discurso.

6. Gradación tiene que ver con que muchas de las características anteriores se manifiestan en mayor o menor grado.

La autora clasifica este tipo de unidades en dos grupos. Por un lado están las unidades que no son enunciados, puesto que "necesitan combinarse con otros signos lingüísticos y equivalen a sintagmas" (Corpas, 1996: 51), como son las colocaciones y las locuciones. Por otro, las expresiones que "están fijadas en el habla y por constituir actos de habla realizados por enunciados completos dependientes o no de una situación específica" (Corpas, 1996: 51), a las que denomina enunciados fraseológicos. Durante el etiquetado del corpus, se encontraron tanto locuciones como enunciados fraseológicos, los cuales se explican con detenimiento a continuación.

5.2.1 Locuciones

Las locuciones son expresiones formadas por más de dos palabras que se combinan sintácticamente para dar lugar a un significado complejo, es decir, el sentido de la locución no puede entenderse como la suma de los significados de sus elementos constituyentes (cfr. *Nueva gramática de la lengua español. Manual*). Las locuciones poseen dentro de la oración la función de una clase de palabra; por lo que, las clases de locuciones pueden clasificarse en gramaticales y léxicas. Las gramaticales se refieren a las locuciones conjuntivas y prepositivas, pues se encargan de unir partes

de la oración. Las locuciones léxicas son aquellas que pueden sustituir un sustantivo, adverbio, adjetivo, etc.

5.2.1.1 Locuciones nominales

Las locuciones nominales sustituyen sustantivos. Pueden tener la función sintáctica de sujeto, objeto directo y objeto de término preposicional (cfr. Carballo, 2015: 126). Las locuciones nominales "están formadas por sintagmas nominales de diversa complejidad". Por ejemplo: *golpe bajo*, *conejillo de indias*.

5.2.1.2 Locuciones adjetivas

Las locuciones adjetivas pueden ser sustituidas por un adjetivo debido a que funcionan en la oración de igual manera: como predicación o atribución; es decir, adyacentes al sustantivo o después de un verbo copulativo. Su núcleo puede ser un adjetivo o un participio (cfr. Carballo, 2015: 128). Algunos casos son *a toda madre, del diablo*, *de huevos*, *de la chingada*.

5.2.1.3 Locuciones adverbiales

Pueden funcionar como adyacentes circunstanciales de verbos, o de toda una oración, al igual que las proposiciones de carácter adverbial, y expresar circunstancias de tiempo modo y cantidad (cfr. Carballo, 2015: 131). Este tipo de locuciones generalmente están formados por sintagmas prepositivos. Sus valores semánticos pueden ser de modo, tiempo y espacio, por lo que este tipo de expresiones pueden funcionar como complementos circunstanciales. Por ejemplo: *a gusto*, *sin remedio*, *a favor*, *de balde*, *en vano.*

5.2.1.4 Locuciones verbales

Este tipo de locuciones expresan estados o procesos y, al igual que el verbo, tienen flexión gramatical de persona, número y tiempo (cfr. Carballo, 2015: 129). Ejemplos:

echar de menos [algo o alguien], *revolver el estómago, valer la pena*, [*alguien*] *me cae bien/mal*.

5.2.1.5 Locuciones interjectivas

Estas locuciones son expresiones de dos o más palabras acuñadas o asimiladas como interjecciones. Por ejemplo: *¡Maldita sea!*, *¡Ah huevo!*, *¡Qué poca madre!*

5.2.2 Enunciados fraseológicos

Dentro de los enunciados fraseológicos, Corpas (1996) realiza una distinción entre las paremias y las fórmulas rutinarias. En las paremias se engloban los refranes (aforismos, sentencias, adagios, etc.), las citas, los lugares comunes, los eslóganes (135-136). En las fórmulas rutinarias se engloban los saludos, despedidas, peticiones, y todas aquellas expresiones que se utilizan en contextos sociales, discursivos particulares y rutinarios, muchos de ellos son fórmulas de cortesía. Las fórmulas rutinarias habían sido tratadas en parte por Casares (1992[1950]). En su clasificación el autor incluye los timos o expresiones de carácter efímero como: *Que te crees tú eso*, *A ver si vas a poder*, *No hay derecho*.

Las fórmulas rutinarias se caracterizan por los siguientes rasgos: "su repetición, dependencia de la situación, su carácter predecible, su (mono-) funcionalidad pragmática y su normatividad interindividual" (Corpas, 1996: 171-172). Dentro de ellas se encuentran las fórmulas psicosociales las cuales son "facilitadores del desarrollo normal de la interacción social, o bien, funciones de expresión del estado mental y los sentimientos del emisor" (192). Algunas fórmulas que expresan portadoras de los sentimientos y disposición del hablante expresan aprobación o rechazo. Por ejemplo:

(13) *¡Qué…ni qué ocho cuartos!*
(14) *A otro perro con ese hueso.*

(15) *No faltaba más.*

A lo largo de este capítulo, se ha discutido sobre las diferentes clases de palabras y unidades fraseológicas, que, de acuerdo con distintas clasificaciones semánticas proporcionan alguna carga positiva o negativa. Como se ha mencionado en el apartado de los diccionarios, en el capítulo de minería de opiniones, una clasificación de este tipo no agota el problema; sin embargo, es una guía para realizar una distinción.

6 CORPUS Y METODOLOGÍA

En este capítulo se explicará la metodología para la recolección del corpus, así como sus características. También se hablará sobre cómo se realizó el etiquetado del mismo y con base en qué criterios.

6.1 Características de *Twitter*

Las publicaciones de los usuarios en *Twitter* son pequeños textos que no deben exceder los 140 caracteres. Por la brevedad e inmediatez, muchos contienen errores de ortografía. En otras ocasiones se trata de simular el lenguaje hablado, y se incurre deliberadamente en dichas faltas: se utilizan mayúsculas para gritar, repetición de grafías para enfatizar. Debido a que la comunicación no se realiza cara a cara, las personas suelen incluir emoticones para mostrar sus emociones con respecto a un tema, o minimizar el impacto de una opinión, dejando entrever que no se espera ser descortés. Los emoticones son una secuencia de caracteres que muestran distintas expresiones faciales, por ejemplo: :), :(.

Además de que el usuario escribe algo interesante para él, puede compartir lo que algún otro individuo o institución haya dicho, esto es, realizar un *retweet* y se distingue porque al inicio contiene las letras 'RT', o bien, si se desea, es factible etiquetar a otro usuario dentro de una publicación, esto puede notarse porque antes del nombre de usuario aparece una @. Se puede hablar de cualquier tema; sin embargo, si se quiere opinar sobre un tema del que todo el mundo está hablando (*trending topic*), se utilizan los *hashtags* (#), esto permite saber lo que los demás opinan acerca del tema y apoyarlo o censurarlo con el fin de que otros lo vean. Los *tweets* también pueden contener imágenes, *gifs*, videos, y urls.

6.2 Recolección del corpus

Se realizó un corpus balanceado de *Twitter* para esta investigación. Se consideró esta red social debido a su naturaleza oral/escrita, ya que en ella la mayoría de los usuarios utiliza un registro informal, esto es, escribe de manera espontánea y poco cuidada (la mayoría de las veces), por lo que se espera encontrar más marcado el rol de género. Además, las pruebas lingüísticas en casos de perfilamiento de autor suelen ser de poca extensión y los *tweets* son textos muy pequeños.

A pesar de que existe una API (interfaz de programación de aplicaciones) de *Twitter* que permite descargar de manera rápida varios miles de *tweets* con algún costo, la recolección del corpus se realizó de forma manual con el fin de garantizar que la información de cada participante fuera verídica y que no se tratara de un bot o perfil falso.

Participaron 40 individuos de esta red social: 20 hombres y 20 mujeres, de 18 a 28 años, con escolaridad superior y habitantes de la Ciudad de México. Se recolectaron los 50 *tweets* más recientes de cada usuario, pero previos a que se les solicitara su apoyo en este experimento para no condicionar sus escritos. No se tomaron en cuenta ni citas, ni *tweets* completamente en otro idioma, ni *retweets* que no contengan producción escrita del usuario, ni sólo imágenes, ni únicamente emoticones. Sin embargo, cuando los *tweets* poseían, además de texto elaborado por la persona, emoticones, etiquetas, hashtags, url y anglicismos, se conservaban en el corpus sin modificación alguna. Se obtuvo un total de 2,000 *tweets* (1,000 por sexo). La cantidad de usuarios y *tweets* es representativa y manual debido a la dificultad que implica identificar una opinión (cfr. cap. 3); por lo que en primer lugar se debía realizar un estudio lingüístico para que un futuro se pueda realizar automáticamente.

Todos los participantes debían ser de la misma región geográfica: la Ciudad de México, pensando que podrían existir variaciones léxicas, sintácticas o morfológicas si pertenecían a distintas partes de la República Mexicana, lo cual podía influir en los resultados, ya que al ser un corpus tan pequeño ciertos usos podrían ser registrados por un sólo sexo. Cabe destacar que el anteriormente llamado Distrito Federal, posee muy diversas colonias tales como Tepito y Santa Fe, donde se pueden registrar dialectos muy dispares, uno considerado de la clase social baja y el otro, de alta. Esto también fue tomado en cuenta. Para disminuir la variación dialectal de la muestra, se pensó que antes que el estrato social, la educación era uno de los factores que disminuía la dispersión.

Los intervalos de edad no son estrictos (cfr. Areiza, 2012:41), no obstante, el ser humano comienza a considerarse que está en su etapa adulta a partir de los veintidós. Sus hábitos lingüísticos cambian debido a que entra en el mercado laboral y debe usar la modalidad prestigiosa de la lengua o el registro estandarizado. A pesar de esto, se consideraron personas desde 18 años, puesto que es la edad en que el mexicano promedio entra a la universidad y esta te exige utilizar un uso estándar de la lengua para desarrollarte en el ámbito, debe mantener un prestigio y una imagen ante la sociedad y ante sus colegas. Todos los participantes cursaron o están cursando una licenciatura.

Antes del etiquetado del corpus, no se efectuó pre-procesamiento alguno. Se conservaron los emoticones debido a que podrían ayudar a desambiguar la polaridad del *tweet*; sin embargo, el propósito de esta investigación no es saber qué sexo utiliza más emoticones, sólo sirvieron como elemento de apoyo. Por otro lado, las etiquetas y *hashtags* tampoco se eliminaron porque en muchos casos, las etiquetas aportan información sobre el sujeto o empresa de la que se opina, y los *hashtags* sobre el tema

y la polaridad del *tweet*. Se pretende que este corpus pueda ser ocupado para estudios ulteriores, por lo que, los errores de ortografía tampoco fueron corregidos, pensando que las mayúsculas y la repetición de caracteres para intensificación puedan ser estudiados como otros factores que sirvan para el perfilamiento, además emoticiones, etiquetas, url, *hashtags*.

6.3 Etiquetado del corpus

Después de la recolección del corpus, se llevó a cabo una separación de *tweets* en "opiniones" y "no opiniones". Se consideraron como "opiniones" aquellos fragmentos que tuvieran las características de una opinión regular u comparativa de acuerdo con las definiciones mencionadas en el capítulo sobre minería de opiniones.[5] Por otro lado, las "no opiniones" son aquellas que no contienen alguno de los elementos característicos de una opinión regular o comparativa o ninguno; el anotador tiene que inferirlos con su conocimiento de mundo o simplemente porque hay indicios de ironía, tema que no entra dentro de los propósitos de la presente investigación. El fragmento (16) ejemplifica una "no opinión".

> (16) El 24 de mayo es el deadline para participar en el proyecto bartkira!!!
> https://i.chzbgr.com/maxW500/7165847808/h62DC7CF9/...#Bartkira

La opinión regular debe contener por lo menos 4 elementos principales: objeto u entidad de la que se opina, opinador, ancla de texto y polaridad, de acuerdo con la propuesta de Wiebe *et al.* (2005). A continuación, se muestra uno de los casos extraídos del corpus y cómo se clasificó cada una de sus partes:

[5] La cantidad de opiniones se detallan en el capítulo de análisis léxico-estadístico.

(17) Breaking bad es una joya de serie… desde la forma hasta el fondo.

Entidad de la que se opina	Opinador	Ancla de texto	Polaridad
Breaking bad	hombre	joya de serie	positivo

Tabla 1. Ejemplo de etiquetado de opiniones regulares

A diferencia de las opiniones regulares, en las comparativas se estiman las diferencias o semejanzas entre dos o más entidades, por lo que estas cuentan con: entidades que se comparan (E1, E2), opinador y aspectos que se comparan, para los cuales se mantuvo el nombre de ancla de texto. En este tipo de opiniones, no existe como tal una polaridad, puesto que no hay una valoración sobre las entidades sino una graduación; sin embargo, para fines prácticos uno puede asignar una opinión positiva por la entidad que prefiere el opinador (cfr. Liu, 2012: 115). En este caso, se prefirió designar una polaridad con respecto a lo que se decía sobre la entidad 1, es decir, si se decía que la entidad 1 era mejor que la entidad 2, la polaridad era positiva y si se decía que la entidad 1 era peor que la entidad 2, la polaridad era negativa. Ahora bien, más abajo se presenta un enunciado etiquetado conforme a los componentes de las opiniones comparativas:

(18) Honestamente no hay mejor voz femenina mexicana que la de @ANAGABRIELRL para cantar el Cielito Lindo

E1	E2	Opinador	Ancla de texto	Polaridad
Ana Gabriel	El resto de las cantantes femeninas mexicanas	Hombre	no hay mejor voz femenina mexicana	positivo

Tabla 2. Ejemplo de etiquetado de opiniones comparativas

Además de las casillas anteriores, a las opiniones comparativas se le agregó una más: la clase de opinión comparativa graduable. Las opiniones comparativas graduables pueden ser: graduables no iguales (del tipo *mayor* o *menor que*), iguales (del tipo *igual que*) y superlativas (*el mejor*/*el peor*). No se consideraron las opiniones comparativas no graduables pues estas solo mencionan qué aspectos comparten o diferencian a las entidades, o si estas son similares o diferentes, pero no se dice que entidad prefiere un opinador, por lo cual, aunque pueden expresar una opinión sobre una entidad esta puede ser difícil de verse.

En los casos previos se puede distinguir de manera sencilla si el texto es una opinión o no, y si este es una opinión, cuáles son las partes que lo conforman. No obstante, se hallaron casos en los que era complicado determinar su polaridad. Obsérvese lo siguiente:

(19) Soy una **romántica** *sin remedio* [Mujer]

(20) ***Me asusta*** lo mucho que me gusta Gossip Girl. ***Maldita sea.*** [Mujer]

(21) A veces me llegan canciones bien simples, sí es un poco ***Guilty Pleasure*** Coldplay - A (Sky Full Of Stars https://youtu.be/VPRjCeoBqrI [Hombre]

(22) ***¿Qué tan culposo es el gusto*** por... ¡Salón Victoria!?? Sí, los tengo en mi iTunes, y qué? :) [Mujer]

(23) **Ando bien clavas** con una canción *de lo más marica* [Hombre]

A estos ejemplos se les nombró como bipolares. Parecen un poco contradictorios, mas implican una disociación del sujeto enunciante en un sujeto evaluador (a partir de ciertos criterios más o menos objetivados) y un sujeto gustador. (cfr. Kerbrat-Orecchioni, 1986: 125). Esto quiere decir que quien tuiteó el ejemplo (21 y 22) en realidad gusta de la canción de Coldplay y la banda Salón

Victoria, pero contienen algunos aspectos que la sociedad o ellos mismos consideran desagradables.

Algunos *tweets* cuentan con más de una opinión sobre entidades diferentes, por lo que fueron divididos.

(25) Me **cagan** *los veganos*, pero, tengo *un hijo vegano* que **no** me **caga**.

a. Me **cagan** *los veganos*

b. tengo *un hijo vegano* que **no** me **caga**.

Entonces los elementos etiquetados como opiniones fueron analizados como se explica a continuación. Primero, se realizó una clasificación temática de las opiniones. Segundo, el ancla de texto fue clasificado de acuerdo con la clase de palabra a la que pertenecía, como se muestra en la tabla 3, asimismo, las diferentes locuciones que reflejaran alguna valoración (locuciones adverbiales, verbales, interjectivas, adjetivas y sustantivas); tal es el caso de 'valer la pena'.

Tweet	opinión	Tipo de opinión	Tipo de opinión comparativa	Tema	E1	E2	Opinador	Ancla de texto	Polaridad	Sustantivo	Adjetivo
Breaking bad es una joya de serie	opinión	Regular		Entretenimiento	Breaking bad		Hombre	Una joya de serie	Positivo	Joya	
La explosión volcánica más hermosa del siglo https://shar.es/1pdxZO vía @culturacolectiv	opinión	comparativa	Superlativa	Naturaleza	la explosión volcánica	El resto de las explosiones volcánicas	Mujer	la más hermosa del siglo	Positivo		Hermosa
El 24 de mayo es el deadline para participar en el proyecto bartkira!!! https://i.chzbgr.com/maxW500/7165847808/h62DC7CF9/… #Bartkira	no opinión										

Tabla 3. Ejemplo de etiquetado del elemento subjetivo

El etiquetado del enunciado de opinión también se realizó de forma manual y sin ayuda de un diccionario de palabras de sentimientos, debido a que este te proporciona palabras previamente consideradas como positivas o negativas (Ortíz *et al.*, 2010; Redondo *et al.*, 2007) o de acuerdo con la probabilidad de que pertenezcan a una emoción (Rangel *et al.*, 2014), y por lo cual, al etiquetar automáticamente pueden hacer falta algunas de ellas que no aparezcan en el lexicón y sí en el corpus, o en caso de que sí estén presentes en el corpus se debe tomar en cuenta que las palabras funcionan de acuerdo con su contexto. Debido a su carácter personal e informal, en el corpus aparecieron varias palabras y expresiones altisonantes. Se documentó la locución verbal 'me caga algo' (ejemplo 25), variante de 'me cago en algo', cuyo significado en el DRAE es "expresar desprecio o rechazo hacia esa persona o cosa." [6]

Se etiquetaron los adjetivos que contienen una evaluación sobre alguna entidad (aunque no se realizó una distinción entre los diferentes tipos), a estos adjetivos Kerbrat Orecchioni (1986) los denomina evaluativos axiológicos y adjetivos afectivos. Los primeros realizan una evaluación entre lo 'bueno' y lo 'malo'; mientras que los segundos, "consideran una propiedad del objeto al que determinan" y la reacción emocional del sujeto frente al mismo (cfr. Kerbrat, 1986). Esta clasificación corresponde con un criterio semántico de clasificación de los adjetivos, es decir, adjetivos evaluativos, los cuales "incluyen los aspectos de la realidad, humana y no humana, que los seres humanos consideran susceptibles de valoración" (Bosque y Demonte, 1999), y adjetivos de habilidades y predisposiciones humanas, los cuales incluyen aptitudes emocionales (*sensible*, *amable*, *cordial*, *cariñoso*), intelectuales

[6] Real Academia Española. (2001). *Diccionario de la lengua española* (22.aed.). Consultado en <http://www.rae.es/rae.html>.

(*inteligente*, *capaz*, *sabio*), y pasiones y disposiciones primordiales (*nervioso*, *agresivo*)" (Bosque y Demonte, 1999).

Se clasificaron sustantivos evaluativos, muchos de ellos derivados de verbos y adjetivos, así como adverbios. Si en algún momento se encontraba una trasposición de categoría, por ejemplo, un cambio de adjetivo a sustantivo, sólo se consideraba cuando la metábasis era lexicalizada, es decir, "el elemento sustantivado se incorpora al léxico de la lengua como sustantivo" (Gualda, 1989: 19); para ello se consultó el DRAE.

En algunas ocasiones se encontraron participios y gerundios fuera de perífrasis verbales (las cuales fueron marcadas como verbos); los primeros fueron encasillados como adjetivos y los segundos, como verbos. Ya que los participios pasados pueden funcionar como adjetivos, a menos, claro, que conserven el significado de proceso propio del verbo, en cuyo caso deben considerarse como verbos (Fages, 2005).

Visto que el adjetivo es la clase de palabra comúnmente reconocida por su función valorativa, se analizó la función sintáctica y el uso de modificadores con el objeto de observar diferencias en cuanto a la preferencia de los *twitteros*.

Se notó que existían otras frases diferentes a las locuciones que también proporcionaban un juicio, conocidos como enunciados fraseológicos, expresiones como '¿qué pedo con…?'. Este tipo frases, aunque tuvieron pocas apariciones en el corpus, al realizar una búsqueda en *Twitter,* se observó que eran muy frecuentes.

Algunas de estas expresiones sí fueron documentadas en algunos diccionarios y textos de fraseología:

(26) *¡Qué … ni qué ocho cuartos/nada!* (Corpas, 1996: 180; Cantera y Blanco, 2007)

Por su poca aparición en el corpus se decidió no realizar ninguna separación interna entre los diferentes tipos de enunciados fraseológicos y se dejaron en una sola gran agrupación.

El paso posterior fue realizar un conteo sobre la cantidad de opiniones y no opiniones; de las opiniones se computó la temática, la polaridad, los tipos de opiniones y la clase de palabra utilizada para opinar. Posteriormente, se elaboró una prueba estadística con el programa computacional *R project*, llamada ji.cuadrada, con el objetivo de medir la dependencia de las variables.

7 ANÁLISIS LÉXICO ESTADÍSTICOS

En este capítulo se presentarán los resultados obtenidos sobre las etiquetas previamente descritas en al capítulo de corpus y metodología así como la relación entre sexo y la clase de palabra utilizada para opinar. Pero antes, se explicará qué método estadístico se utilizó para ver la dependencia entre las variables.

7.1 Ji cuadrada

A través de métodos de estadística inferencial se pueden obtener conclusiones sobre grandes grupos o eventos con base en una muestra de población. La ji.cuadrada es uno de esos procedimientos, esta prueba estadística puede utilizarse para determinar lo significativo de las diferencias entre dos grupos independientes (cfr. Siegel, 1972) o para ver si los niveles de la variable independiente resultan en diferentes frecuencias de los niveles de la variable dependiente (cfr. Gries, 2013: 179).

Primero, se deben formular dos hipótesis: H0 y H1. Las cuáles se pueden explicar de la siguiente manera:

> H0: es una hipótesis de no efecto o nula, donde las frecuencias de los niveles de la variable dependiente no varían como una función de los niveles de la variable independiente, $\chi^2 = 0$
>
> H1: las frecuencias de los niveles de la variable dependiente varían como una función de los niveles de la variable independiente, $\chi^2 > 0$

Segundo, se realiza la ji.cuadrada con base en las frecuencias obtenidas para cada uno de los grupos con sus respectivas variables como se muestra a continuación:

	Grupo	
Variable	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}
Total	C_1	C_2

Tabla 4. Tabla de contingencias[7]

Entre más cercano a 0 es el resultado de la ji.cuadrada, quiere decir que la distribución es aleatoria y por tanto no depende de uno de los grupos; sin embargo, para aceptar o rechazar la hipótesis nula se deben considerar los grados de libertad. Cuando se tienen dos categorías se debe restar 1 para obtener los grados de libertad, debido a que la frecuencia de una categoría puede variar libremente pero la otra está fija.

Con base en estos dos resultados se proporciona un *p-value* o probabilidad de significación, esto es, la probabilidad de error de aceptar que H1 dados los resultados obtenidos. Según Gries (2013: 29) se pueden encontrar diferentes expresiones para diversos *p-values*:

- $p<0.001$ a veces se refiere a un altamente significante y se indica con ***
- $0.001\leq p< 0.01$ a veces se refiere a que es muy significante y se indica con **
- $0.01\leq p<0.05$ a veces se refiere a que es significante y se indica con *
- $0.05 \leq p < 0.01$ a veces se refiere un marginalmente significativa y se indica con ms, este *p-value* es mayor que el estandar 5% en realidad no es tan significante.

[7] Tabla tomada de Siegel, 1972.

Por tanto, para las pruebas estadísticas que se realizaron, se consideró que si el *p-value* era menor a 0.05 se podía rechazar la hipótesis nula y aceptar la alternativa.

7.2 Resultados

La siguiente tabla muestra la cantidad de opiniones y no opiniones encontradas en el corpus. Recuérdese que en un apartado anterior, se mencionó que algunos *tweets* contenían más de una opinión y fueron separados. Es por ello que la suma de 'opinión' y 'no opinión' no resulta en la cantidad de *tweets* que fueron extraídos en primer lugar (1000).

	Hombres		Mujeres	
Opiniones	686	61.19%	720	65.27%
No opiniones	435	38.80%	383	34.72%
p-value $= 0.05098$				

Tabla 5. Opiniones

En la tabla anterior puede notarse que existe una diferencia del 4.08% entre la cantidad de opiniones femeninas y masculinas, pero el *p-value* de la ji.cuadrada[8] es 0.05098, lo cual indica que el hecho de que un *tweet* sea una opinión o no, no está relacionado con el sexo de su autor.

Las figuras 2 y 3 muestran los temas más recurrentes en el corpus de opiniones. 'Otros' contiene el resto de temas con un porcentaje de aparición menor. En el Apéndice I se puede observar desglosadamente el tema y la cantidad de opiniones por cada uno.

[8] El programa señala una corrección continua de Yates, lo cual indica que el corpus es pequeño.

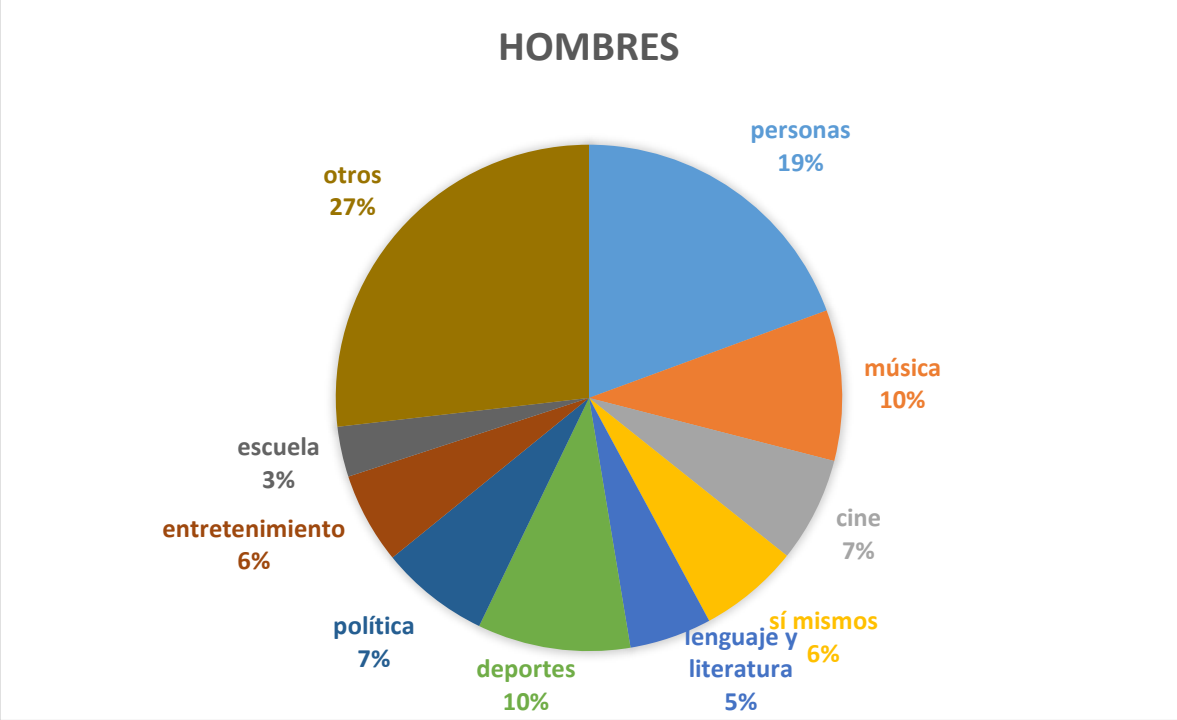

Figura 2. Temas más comunes de las opiniones masculinas

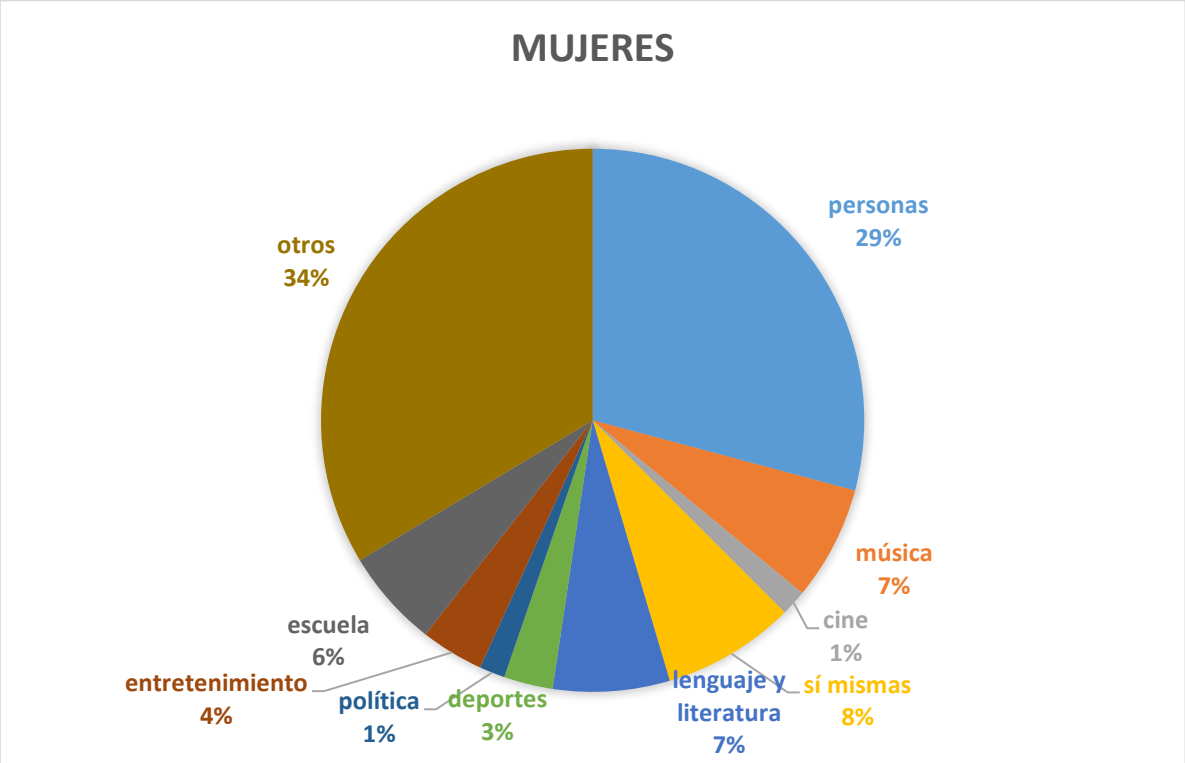

Figura 3. Temas más comunes de las opiniones femeninas

En ambos casos el tema más frecuente fue 'personas', con 19% y 29%, respectivamente. Había más *tweets* masculinos sobre política (7% vs 2%), deportes (10% vs 3%) y cine (7% vs 1%). En la categoría 'otros', las mujeres tienen 34% de sus opiniones y los varones 27%.

La tabla 6 describe el conteo de la polaridad de las opiniones. Como puede notarse, las mujeres realizaron más opiniones positivas que los varones, 3.69% más; mientras que los hombres realizaron más opiniones negativas, 3.92% más. El resultado del *p-value* fue mayor a 0.05 (0.3327), esto quiere decir que la polaridad de la opinión es independiente del sexo.

	Hombres		Mujeres	
Positivo	331	48.25%	374	51.94%
Negativo	347	50.58%	336	46.65%
Bipolar	8	1.16%	10	1.38%
Total	686		720	
p-value= 0.3327				

Tabla 6. Polaridad de las opiniones

La cantidad de los tipos de opinión, regular y comparativa, es mostrada en la tabla 7. De las regulares, 89.16% fueron femeninas y 88.19%, masculinas. De las comparativas, 10.83% fueron femeninas y 11.80%, masculinas. El *p-value* es de 0.3433 por lo tanto no existe dependencia entre sexo y tipo de opinión.

	Hombres		Mujeres	
Opiniones regulares	605	88.19%	642	89.16%
Opiniones comparativas	81	11.80%	78	10.83%
Total	686		720	
p-value = 0.3433				

Tabla 7. Tipos de opiniones

La tabla 8 muestra las diferentes opiniones comparativas graduables que aparecieron en el corpus. No hubo diferencias significativas.

	Hombres		Mujeres	
Superlativo	56	69.13%	60	76.92%
Desigualdad	22	27.15%	17	21.79%
Igualdad	3	3.70%	1	1.28%
Total	81		78	

Tabla 8. Tipos de opiniones comparativas

La tabla 9 y 10 muestran la categoría léxica que contiene la valoración del opinador. Como puede notarse, el total de elementos léxicos (815 y 867) no coincide con la cantidad de opiniones emitidas (686 y 720), ya que en varios casos, para una única entidad se utilizaban dos o más elementos subjetivos. En la tabla 9, aparecen separadas las clases de palabras y las locuciones con el objeto de ver si utilizar una construcción de más de dos palabras tiene alguna significancia. En la tabla 10, las locuciones se sumaron con la categoría a la que corresponden (verbos + locuciones verbales, sustantivos + locuciones nominales, etc.) y los enunciados fraseológicos fueron añadidos a las interjecciones.

	Hombres	%	Mujeres	%
Sustantivos	147	18.03	122	14.07
Adjetivos	433	53.12	407	46.94
Verbos	132	16.19	229	26.41
Interjecciones	16	1.96	23	2.65
Adverbios	21	2.57	12	1.38
Locuciones	37	4.53	51	5.88
Enunciados fraseológicos	29	3.55	23	2.65
Total	815		867	
p- $value = 8.526e\text{-}05$				

Tabla 9. Elemento utilizado para opinar

	Hombres	%	Mujeres	%
Sustantivos	147	18.03	122	14.07
Adjetivos	439	53.86	414	47.75
Verbos	146	17.91	258	29.75
Adverbios	26	3.19	17	1.96
Interjecciones	57	6.99	56	6.45
Total	815		867	
p-value = 6.101e-07				

Tabla 10. Clase de palabra utilizada para opinar

De los resultados anteriores se puede deducir que los adjetivos son la clase de palabra más utilizada para opinar por ambos sexos. Representa la mitad de la muestra aproximadamente. Con respecto a la tabla 10, los adjetivos son utilizados 6.11% más por los hombres y los sustantivos 6.18%. Mientras que los verbos son utilizados más por las mujeres, 11.88% más, en el resto de clases de palabras existe una diferencia del 1% como máximo.

Posteriormente, se realizó una prueba estadística con el objeto de descartar alguna dependencia del elemento utilizado para opinar y el sexo de la persona que emite la opinión. En ambos casos el resultado de la ji.cuadrada mostró que el elemento utilizado para opinar sí es dependiente del sexo. Para la tabla 9 se obtuvo un *p-value* de 8.526e-05 y para la tabla 10, 6.101e-07; sin embargo, en este tipo de *test* no resulta claramente observable qué variables tienen dependencia con el sexo. Ante esta situación Siegel *et al.* (1972) recomienda realizar una separación de la frecuencia de cada unidad y compararla con el resto de elementos, y de esta manera

se puede saber qué variable es dependiente. En este trabajo, los resultados de dependencia solo fueron positivos para dos clases de palabras: los adjetivos y los verbos. En la tabla 11 y 12 se presentan las tablas de contingencias donde se comparan las frecuencias de verbos y adjetivos contra el resto de los elementos opinativos. El *p-value* de los adjetivos fue 0.01289; el de los adjetivos + locuciones adjetivas, de 0.01398; el de los verbos, 4.632e-7, y el de los verbos + locuciones verbales, 1.71e-08; lo cual indica que la relación de dependencia entre los adjetivos es significativa; mientras que la de los verbos es altamente significativa.

	Hombres	Mujeres		Hombres	Mujeres
Adjetivos	433	407	Adjetivos + locuciones adjetivas	439	414
Resto de elementos opinativos	382	460	Resto de elementos opinativos	376	453
p-value = 0.01289			*p-value* = 0.01398		

Tabla 11. Adjetivos contra el resto de elementos de opinión

	Hombres	Mujeres		Hombres	Mujeres
Verbos	132	229	verbos + locuciones verbales	146	258
Resto de elementos opinativos	683	638	Resto de elementos opinativos	669	609
p-value = 4.632e-07			*p-value*= 1.71e-08		

Tabla 12. Verbos contra el resto de elementos de opinión

A continuación, se presentan algunos ejemplos donde se marcan los elementos que se consideron como el elemento opinativo. En (27) y (28) se utilizó la locución interjectiva *¡No mamar!,* mientras en (27) se utiliza en un contexto positivo en (28) se usa de forma negativa. En todos los casos que apareció esta interjección en las opiniones de mujeres se utilizó de manera negativa (4 veces) y en los hombres fue usada tanto de manera positiva (3 casos) como negativa (5 casos).

(27) **No mames**, qué bueno es el Rey León [Hombre]

(28) @katzehexe @Bucktigamesh Cup caques.... **no mamen** mi inglés es **pésimo**... nahh tengo 3 años para aprender italiano :3 [Mujer]

Algunos elementos que llamaron la atención fueron, en primer lugar, las groserías: ambos sexos tuvieron una frecuencia muy similar pese a que se ha dicho que las mujeres dicen menos palabras de este tipo (Lakoff, 1975; Schwartz). La palabra *pendejo* obtuvo el mismo número de frecuencias (11), mientras *pinche* alcanzó 15 apariciones por los hombres y 13 por las mujeres.

En el terreno axiológico, el adjetivo *bueno* (49 vs 14), *chido* (11 vs 5) y *malo* (10 vs 6) fue preferido por los varones; mientras que las féminas se inclinaron por

bonito (19 vs 4) y *hermoso* (12 vs 4). Cuando las palabras se refieren a sentimientos las mujeres los emplearon en mayor cantidad *amor* (7 vs 1), *miedo* (7 vs 2), *felicidad* (4 vs 0), *triste* (8 vs 7), *feliz* (8 vs 5). La suma de estas palabras de sentimientos da 34 para el sexo femenino y 15 para el masculino.

A continuación, se hablará sobre las clases de palabras en las cuales se encontraron diferencias significativas de uso de acuerdo con la prueba estadística: los verbos y los adjetivos. Se decidió que los adjetivos debían estudiarse con mayor detalle ya que es la clase de palabra más común para expresar una valoración y los verbos porque presentaron un *p-value* altamente significativo.

7.3 Verbos

Los resultados de la ji.cuadrada mostraron que existe un alto grado de significancia para rechazar la hipótesis nula y aceptar la alternativa. La variable "verbo" no es independiente del sexo.

Posteriormente, se observó que los verbos psicológicos o verbos de estado anímico, es decir, aquellos que "denotan estados emocionales como *temer*, *gustar* y *molestar* y que implican dos argumentos: un experimentante, que refiere a la persona que tiene la emoción indicada por el verbo, y un tema, que refiere a la entidad relacionada con la emoción" (Vroon, 2006:1) eran los verbos más usados. Se realizó una separación entre los verbos psicológicos y el resto y se comprobó que las mujeres usaron más verbos de estado anímico. En la figura 4 se aprecia que ellas lo utilizaron más del doble que ellos.

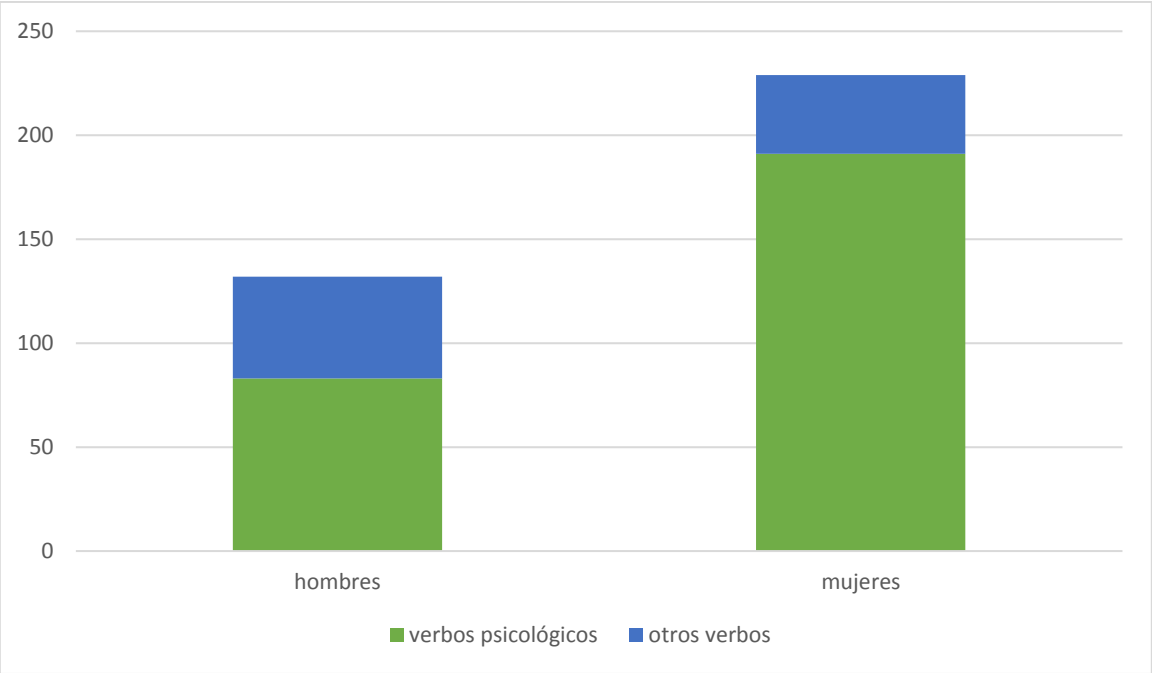

Figura 4. Tipos de verbos

A continuación, se presentan los verbos psicológicos más utilizados.

	Hombre	Mujer
amar	6	39
gustar	12	26
querer	11	22
adorar	0	12
extrañar	8	16
odiar	4	20

Tabla 13. Los verbos psicológicos más utilizados

También se realizó una ji.cuadrada para la semántica de los verbos. Se obtuvo un *p-value* de 2.006e-05. Este resultado nos indica que hay una alta significancia de dependencia entre el sexo y el tipo de verbos utilizados.

(29) a. Te *Amo*, Pumas, porque eres el único equipo que tiene a Gustavo Adolfo Rodríguez en la cantera.

b. Sí, se llama Gustavo Adolfo. Sí, lo *amo*.

(30) No sé sí entienden cuánto *amo* las danzas polinesias. Como sea, seguiré bailando hasta tener mi Ballet Polinesio así como Lourdes. #HeDicho

(31) Excelente lectura y excelente serie, *recomiendo* ambas. #TrueDetective #erratanaturae http://instagram.com/p/xfyK2rAw3M/ [Hombre]

(32) *Recomiendo* a @coeur_illimite, va empezando pero sé de varios que les gustarán sus *tweets* =)

Los ejemplos (28) y (29) fueron extraídos del corpus y contienen una valoración en la que para expresarla se utiliza un verbo psicológico, ambos con el verbo *amar*. En el primer caso se expresa una opinión positiva sobre el equipo de fútbol y otra sobre el futbolista. En el segundo sobre las danzas. En los ejemplos (29a) y (30) el estímulo tiene menor animacidad mientras que en (29b) el estímulo es más animado. En los ejemplos (31) y (32) se utiliza un verbo de decir, el verbo recomendar en gerundio.

La figura 5 muestra los verbos sin las locuciones verbales. Algunas de las locuciones encontradas se podrían considerar como psicológicas: *tener hasta la madre* (H= 0, M =1), *dar lo mismo* (H = 0, M= 1), *tomarse muy a pecho* (H = 0, M= 1), *caer bien/mal* (H = 3, M = 6) y *cagarse* [en] (H = 3, M = 8). En cualquier caso, las chicas se sirven de ellos con mayor frecuencia, pues del total de verbos + locuciones, 208 serían los femeninos y 89 los masculinos.

7.4 Adjetivos

El adjetivo es definido como una clase de palabra que modifica al sustantivo o predica de él aportándole propiedades o cualidades. Los adjetivos admiten modificadores que pueden incrementar o disminuir la intensidad de una opinión, dichos modificadores son adverbios de grado como *muy*, *poco*, *bastante*, *harto*, *qué*, etc. y pueden formar parte de construcciones comparativas o de superlativo.

En la tabla 14 se presentan los modificadores adverbiales y sus frecuencias de uso para ambos sexos. Se incluyen algunos sufijos al análisis puesto que aunque son recursos morfológicos también añaden intensificación a la opinión. En el corpus, se encontró que las mujeres realizan mayor modificación de los adjetivos: 33.16% frente a 26.09%, esto es, ellas prefirieron modificar sus opiniones, atenuarlas o intensificarlas.

	Hombres	Mujeres
Algo	0	1
Bastante	0	1
Bien	10	14
Completo	0	1
Demasiado	1	3
Especialmente	1	0
Igual	1	0
-ísimo	5	6
Lo(la, el) más	19	12
Más	18	15

mayor	0	1
Muy	12	20
Medio	3	0
-ota	1	0
[No hay] nada[como](más)	0	6
Profundamente	1	0
Puras	1	0
Qué	13	20
Re-	1	1
Realmente	2	0
Sobremanera	0	1
Sumamente	0	1
Súper	3	2
Tan	19	25
Tan más	1	0
Todo(a/s)	1	2
Tremendamente	0	1
Un poco	0	1
no	9	12
Total	122	147

Tabla 14. Modificadores de los adjetivos

En (33) y (34) se muestran dos ejemplos de intensificación. En el primero se utiliza el prefijo *re-* para intensificar la opinión y en el segundo, el adverbio *medio* para atenuar el juicio de valor.

(33) Esa Laura Bozzo está ***re****loca*.

(34) La banda del Live es **medio** *naquita.*

Con respecto a la función del adjetivo (Tabla 15), los hombres emplearon más adjetivos en la función atributiva (50.57%), y las mujeres más en la función predicativa (45.94%). También se realizó una ji.cuadrada. El *p-value* para las funciones fue 0.5241. De ahí se infiere que no existe dependencia entre el sexo y la función en la que se utiliza el adjetivo.

	Hombres		**Mujeres**	
Atribución	219	50.57%	172	42.26%
Predicación	159	36.72%	187	45.94%
Nominalización	42	9.69%	38	9.33%
Vocativo	13	3.00%	10	2.45%
Total	433		407	

Tabla 15. Función del adjetivo

En los ejemplos (33) y (34) también se mostró que los adjetivos están empleados en la función predicativa pues se utiliza un verbo copulativo para unir al verbo con su adjetivo. En los siguientes, se ocupa la función atributiva, *sexy* modifica directamente a *morra* y *pinche* al cantante.

(35) Yolandi Visser es la morra más sexy de la vida.

(36) ayyyy ese pinche #SamSmith ya me tiene hasta la madre

Cuando se realiza una nominalización, se puede resaltar la parte de un todo como en (37) o denotar a la persona por esa característica como en (38).

(37) Lo malo de los besos es que crean adicción

(38) Acepto que siempre he sido una ridícula.

Si se hace una función vocativa del adjetivo se está utilizando la función emotiva de la lengua como en (39).

(39) La neta me dan muchísimo oso las personas que aprovechan los disfraces de jalouin para mostrar su cuerpo y verse "súper sexies". **¡Ridículos!**

7.5 Discusión

Los resultados aquí obtenidos contrastan con la conclusión de Rivera (2015), quien dijo que el adjetivo es un buen indicador para identificar sexo. El adjetivo por sí solo no es la mejor clase de palabra para distinguir el género de la persona, por lo menos no en los *tweets* y no para opinar, pues estos son palabras que expresan una valoración por excelencia y son usadas con mayor frecuencia por ambos sexos. O quizás porque haya habido una variación diacrónica de las preferencias. La prueba estadística reveló que la dependencia entre el uso de adjetivos y el sexo es significativa; en cambio, el uso de verbos es altamente significativa.

Los procesos psicológicos, los cuales se expresan con verbos de estado anímico o de sentimientos resultaron ser un buen distinguidor del género femenino para este trabajo. También fueron encontrados por Schwartz (2013) en un corpus de estados de *Facebook*. Este tipo de verbos implican que el hablante se manifiesta de manera explícita como el evaluador del objeto.

Por otro lado, los hombres resultaron difíciles de caracterizar, como ha señalado Janssen y Murachver (2004), pues ellos utilizaron todo tipo de palabras para opinar sin darle preferencia a una clase en particular. Sin embargo, mostraron preferencia en el terreno axiológico por los adjetivos *bueno*, *chido* y *malo*.

En los trabajos de Patra *et al.* (2013), Weren *et al.* (2014) y Pimas (2016) utilizaron palabras de sentimientos como uno de los rasgos para perfilamiento automático de sexo; sin embargo, no encontraron que dichos vocablos fueran de mucha ayuda para clasificar los textos. En esta tesis se encontró que las palabras de sentimientos si pueden ayudar a perfilar sexo, no si se toma en cuenta la polaridad de estas, sino la clase a la que pertenecen; pues como se vio en los resultados aquí obtenidos aunque las mujeres fueron más positivas en sus opiniones (51.99% vs 48.25%) y los hombres más negativos (50.58% vs 46.65%), la prueba estadística mostró que estas diferencias no son significativas.

En cuanto a las groserías señaladas por Lakoff (1975) y Schwartz (2013) como una marca del discurso masculino, en realidad aquí no hubo mucha diferencia en el uso. Por otro lado, sobre algunas palabras de que denotan sentimientos como miedo, felicidad y amor fueron preferidas por las mujeres, similar a resultados obtenidos por Schwartz (2013).

8 RESUMEN, CONCLUSIONES Y TRABAJO FUTURO

Durante esta disertación se habló sobre los temas que se tocan a continuación.

8.1 Resumen

En el capítulo dos se definió "lingüística forense" como la conexión entre la lengua y el derecho, porque utiliza el conocimiento lingüístico para mejorar la interacción entre los implicados del proceso judicial y para analizar pruebas lingüísticas clave para un problema legal. Una de las tareas que se llevan a cabo en el área es el perfilamiento de autor, este consiste en encontrar en el texto la mayor cantidad de elementos lingüísticos asociados con un sexo, un grupo etario, una clase social, un nivel de escolaridad y región geográfica, con el objeto de proporcionar sospechosos a la policía cuando no se tiene ninguno o la cantidad es muy grande. Es posible reconocer marcas de grupos sociales gracias a que estos influencian el estilo de un autor.

En el capítulo 3 se definió minería de opiniones como un área del Procesamiento del Lenguaje Natural que tiene por objeto de estudio identificar de manera computacional la información subjetiva dentro de un texto y extraerla. Como información subjetiva se consideran los estados privados, y las opiniones son un tipo dentro de ellos. Las opiniones pueden ser regulares o comparativas y deben poseer los siguientes elementos: opinador, ancla de texto, objeto del que se opina y polaridad.

En el capítulo 4 se comentaron algunos de los trabajos que concentran sus esfuerzos en encontrar las preferencias lingüísticas para distinguir hombres y mujeres, principalmente para el inglés. Los artículos ahí expuestos utilizaron enfoques sociolingüísticos, de perfilamiento automático y minería de opiniones. Los autores concluyen que los resultados obtenidos en diversos corpus apoyan la idea de las

mujeres son más emocionales y personales, mientras los hombres son más informativos; sin embargo, ninguno estudia el campo de la subjetividad a nivel enunciado en relación con todos los elementos léxicos que se pueden utilizar, tanto palabras como expresiones. Este trabajo se centró en las opiniones por la imposibilidad de trabajar con todos los estados privados.

En el capítulo 5 se describieron las categorías léxicas y frases que expresan de alguna manera una valoración sobre una entidad: sustantivos, adjetivos, adverbios, interjecciones, locuciones y enunciados fraseológicos. Los adjetivos han sido los más descritos como elementos opinativos por el hecho de que proporcionan cualidades a los sustantivos. Las clasificaciones semánticas mostraron que hay adjetivos calificativos más subjetivos porque tienen una valoración axiológica o una carga afectiva. Por su parte, los verbos pueden evaluar el proceso denotado por el mismo o un objeto del proceso: una cosa, un individuo o un hecho. Los verbos psicológicos evalúan un elemento del fenómeno y a la vez expresan una afectación en el experimentante.

En el capítulo 6 se describió detalladamente la conformación del corpus de *Twitter*, así como las características más importantes de esta red social. Se eligió esta red social por sus rasgos de oralidad/escritura, las cuales permiten saber cómo se expresa cada sexo de manera más o menos espontánea. Posteriormente, se habló sobre el etiquetado del corpus, los problemas encontrados y cómo estos fueron resueltos.

8.2 Conclusiones

En el capítulo 7 se mostraron los resultados obtenidos: hombres y mujeres opinan en proporciones similares en *Twitter.* Se encontraron algunas diferencias en los temas de los cuales opinan; sin embargo, la polaridad y los tipos de opinión que usan no es un rasgo distintivo. Esto quiere decir que tanto hombres como mujeres opinan en proporciones similares, comparan las características de productos y servicios también en proporciones similares, ni se puede afirmar que son más positivas o negativas sobre sus opiniones.

Los adjetivos evaluativos son la clase de palabra más usada comúnmente para opinar (se utilizó casi el 50% de las veces en ambos casos) y la prueba estadística mostró que hay significancia para aceptar dependencia. En el corpus, se encontró que las mujeres realizan mayor modificación de los adjetivos: 33.16% frente a 26.09%, esto es, ellas prefirieron modificar sus opiniones, atenuarlas o intensificarlas. En cuanto a la función del adjetivo, los hombres emplearon más la función atributiva (50.57%), y las mujeres más la función predicativa (45.94%).

Por otro lado, los verbos mostraron tener un alto grado de significancia para distinguir el autor de una opinión:

- las mujeres utilizaron los verbos 26.41% (29.75% si se cuentan las locuciones verbales)
- los hombres sólo 16.19% (17.91% si se cuentan las locuciones verbales)
- las mujeres prefirieron utilizar los verbos psicológicos más del doble que los varones (191 frente a 83).

8.3 Trabajo futuro

Para trabajo futuro, se podrá realizar una distinción semántica entre los tipos de adjetivos de acuerdo con la clasificación de Bosque y Demonte, así como de los tipos de sustantivos y los otros tipos de verbos con el objetivo de realizar un estudio con mayor profundidad y en un corpus de mayor tamaño. Se podrá efectuar un análisis de correspondencias para ver si las conclusiones aquí obtenidas son similares o no, y si hay discrepancia, investigar por qué ocurre esto.

Sería interesante investigar si los resultados obtenidos en esta tesis también pueden ayudar a diferenciar grupo etario y nivel de escolaridad. Además, se puede explorar la subjetividad del género y grupo etario en otro tipo de estados privados, como los deseos, las creencias, etc.

9 REFERENCIAS

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.

Areiza Londoño, Rafael, Cisneros Estupiñán, Mireya y Tabares Idárraga, Luis E. (2012). *Sociolingüística: enfoques pragmático y variacionista.* Bogotá, Ecoe ediciones.

Argamon, S., Koppel, M., Finec, J., & Shimonib, A. R. (2003). Gender, Genre, and Writing Style in Formal Written Texts. *To appear in Text*, *23*, 3.

Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. *Proceedings of the WASSA*, *12*.

Belletti, A., & Rizzi, L. (1988). Psych-Verbs and Theta-Theory', NLLT6, 291-352. *Belletti2916NLLT1988.*

Berutto, Gaetano. (1974). *La socioligüística.* Bolonia: Nueva Imagen.

Bora, N. N. (2012). Summarizing public opinions in tweets. *International Journal of Computational Linguistics and Applications*, *3*(1), 41-55.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (pp. 1-45). Technical report C-1, the center for research in psychophysiology, University of Florida.

Cantera, J., & Blanco, P. G. (2007). *Diccionario de fraseología española: locuciones, idiotismos, modismos y frases hechas usuales en español [su interpretación].* Abada Editores.

Carballo, M. A. C. (2015). *De la investigación fraseológica a las decisiones fraseográficas: un estudio de interrelaciones.*

Casares, J. (1992[1950]). *Introducción a la lexicografía moderna.* Madrid: C.S.I.C.

Corpas Pastor, Gloria (1996) *Manual de fraseología española*. Madrid: Gredos.

Coulthard, M., & Johnson, A. (2007). *An introduction to forensic linguistics: Language in evidence*. Routledge.

Coulthard, M., & Johnson, A. (2010). *The Routledge handbook of forensic linguistics*. Routledge.

Demonte, V., & Bosque, I. (1999). *Gramática descriptiva de la lengua española*. Espasa Calpe.

Dixon, R.M.W. (1991). *A new approach to English grammar, on semantic principles*. Clarendon press, oxford.

Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (Vol. 6, pp. 417-422).

Fages Gironella, Xavier. (2005) *Gramática para estudiantes*.Barcelona: Laertes didáctica.

Gil Gaya, Samuel (1993[1943]. *Vox: Curso superior de sintaxis española*. Barcelona: Bibliograf.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*, 12.

Gries, S. T. (2013). *Statistics for linguistics with R: a practical introduction*. Walter de Gruyter.

Gualda, M. V. R. (1989). *El nombre: sustantivo y adjetivo*. Arco Libros.

International Association of Forensic Lingüistics (IAFL). (2013). Consultado en: http://www.iafl.org/index.php

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, *10*(4), 439-459.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, *51*, 35-47.

Janssen, A., & Murachver, T. (2004). The Role of Gender in New Zealand Literature Comparisons Across Periods and Styles of Writing. *Journal of Language and Social Psychology*, *23*(2), 180-203.

Jindal, N., & Liu, B. (2006, July). Mining comparative sentences and relations. In *AAAI* (Vol. 22, pp. 1331-1336).

Johnstone, Barbara (1993). Community and contest: Midwestern men and women creating their worlds in conversational storytelling en *Gender and conversational interaction* : 62-80.

Kerbrat-Orecchioni, C. (1986). La enunciación. *De la subjetividad en el lenguaje. Buenos Aires: Hachette*.

Lakoff, R. (1973). Language and woman's place. *Language in society*, *2*(01), 45-79.

Langacker, R. W. (1985). Observations and speculations on subjectivity. *Iconicity in syntax*, *1*(985), 109.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Lorenzo-Dus, N., & Bou-Franch, P. (2003). Gender and politeness: Spanish and British undergraduates' perceptions of appropriate requests. *Género, lenguaje y traducción, Valencia, Universitat de Valencia/Dirección General de la Mujer*, 187-199.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT press.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Mohammad, S. M., & Yang, T. W. (2011). Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis (acl-hlt 2011*(pp. 70-79).

Mohammad, S. M., & Turney, P. D. (2013). *Nrc emotion lexicon*. NRC Technical Report.

Montejo-Raez, A., Martínez-Cámara, E., Martin-Valdivia, M. T., & Urena-Lopez, L. A. (2012, July). Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 3 10). Association for Computational Linguistics.

Nagoshi, J. L., Brzuzy, S., & N Agoshi, C. A. T. (2014). *Gender and sexual identity*. Springer Science.

Olsson, J., & Luchjenbroers, J. (2013). *Forensic linguistics*. A&C Black.

Ortiz, A. M., Pozo, Á. P., & Sánchez, S. T. (2010). Sentitext: sistema de análisis de sentimiento para el español. *Procesamiento del Lenguaje Natural*, *45*, 297 298.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive science*, *11*(3), 341-364.

Patra, B. G., Banerjee, S., Das, D., Saikh, T., & Bandyopadhyay, S. (2013). Automatic Author Profiling Based on Linguistic and Stylistic Features. *Notebook for PAN at CLEF*.

Pimas, O., Rexha, A., Kröll, M., & Kern, R. Profiling microblog authors using concreteness and sentiment.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., & Crystal, D. (1985). *A comprehensive grammar of the English language* (Vol. 397). London: Longman.

Rangel, I. D., Guerra, S. S., & Sidorov, G. (2014). Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein*,*29*(1), 31-46.

Real Academia Española. (2001). *Diccionario de la lengua española* (22.aed.). Consultado en <http://www.rae.es/rae.html>.

----------(2010). *Manual de la nueva gramática de la lengua española*. Madrid. Asociación de Academias de la Lengua Española.

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (affective norms for English words). *Behavior research methods*, *39*(3), 600-605.

Rivera Vidal, M.A. (2015). *Sociolingüística de los adjetivos calificativos en un corpus del español mexicano.* Tesis de maestría, Universidad Nacional Autónoma de México (UNAM).

Samar, R. G., & Shirazizadeh, M. (2010). Gender-preferential Linguistic Elements in Applied Linguistics Research Papers: Partial Evaluation of a Model of Gendered Language.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006, March). Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (Vol. 6, pp. 199-205).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Sedano, L. H. (2006). Un acercamiento a la gramática de los verbos volitivos, de influencia y psicológicos. In *Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística* (pp. 896-916).

Siegel, S., Villalobos, J. A., Ruseil, L. J., & Cruz-López, R. V. (1972). *Estadística no paramétrica aplicada a las ciencias de la conducta*. México, Trillas.

Strapparava, C., & Mihalcea, R. (2007, June). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 70-74). Association for Computational Linguistics.

Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: an Affective Extension of WordNet. In *LREC* (Vol. 4, pp. 1083-1086).

Shuy, R. W. (2005). La aportación de la lingüística al estudio de la intencionalidad criminal. *Lingüística forense, lengua y derecho*, 1000-1030.

Shuy, Roger W. (2014) Lingüistic profiling when there is no known murder suspect. *The language of murder cases: intentionality, predisposition and voluntariness*. Oxford scholarship online.

Traugott, E. C. (1989). On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language*, 31-55.

Turell, M. T. (Ed.). (2005). *Lingüística forense, lengua y derecho: conceptos, métodos y aplicaciones* (Vol. 8). Documenta Universitaria.

Vroon, S. (2006). ¿Le está gustando la música o lo molesta el ruido? Una investigación sobre el aspecto semántico de los verbos psicológicos y su uso con las formas de los pronombres átonos y en las perífrasis de gerundio.

Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., & Wives, L. K. (2014). Examining multiple features for author profiling. *Journal of Information and Data Management*, *5*(3), 266.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, *39*(2-3), 165-210.

Whissel, C. M. (1989). The dictionary of affect in language, Emotion: Theory, Research and Experience: vol. 4, The Measurement of Emotions, R. *Plutchik and H. Kellerman, Eds., New York: Academic*.

10 APÉNDICES

10.1 Apéndice I

Frecuencias de los elementos de opinión y temas de las opiniones

Tabla. Interjecciones

	Hombres	Mujeres
¡Uju!	0	2
¡Chale!	2	4
¡Ay!	0	3
¡Eh!	0	1
¡Ay no!	1	1
¡Yes!	0	1
¡Yuck!	0	1
¡Guácala!	0	1
¡Mmm!	0	2
¡Buu!	0	1
¡Yay!	0	1
¡Ugh!	1	1
¡Dios!	0	1
¡Joder!	0	1
¡Ash!	0	1
¡Enhorabuena!	0	1
¡Diablos!	1	0
¡Carajo!	2	0
¡Ajj!	1	0
¡Pff!	2	0
¡Mierda!	1	0
¡Wuju!	2	0
¡Verga!	1	0
¡Tss!	1	0
¡No!	1	0
Total	16	23

Tabla. Adverbios

	Hombres	Mujeres
Bien	9	4
Mal	4	1
Difícilmente	0	1
Alarmantemente	0	1
Desgraciadamente	0	1
Estúpidamente	0	2
Tanto	0	1
terriblemente	0	1
Demasiado	1	0
Irracionalmente	1	0
Mejor	5	0
Merecidamente	1	0
Orgullosamente	1	0
Total	21	12

Tabla. Locuciones

	Hombres	Mujeres
*locuciones interjectivas		
¡Maldita sea!	0	3
¡Qué poca madre!	1	0
¡No ma(mes/mar)!	8	4
¡Ah huevo!	1	1
¡Puta madre!	1	2
¡Se la mama!	1	0
*locuciones adverbiales		
De acuerdo	1	0
A favor	1	0
A gusto	1	1
En serio	1	0
Sin remedio	0	1
A la verga	1	0
A la mierda	0	1
De malas	0	1
Un carajo	0	1

*locuciones verbales		
Ser (algo) el colmo	0	1
Caer (bien/mal)	3	6
Ser de cajón (algo)	1	0
Valer la pena	1	1
(Re)chingar a tu madre	3	0
Tomarse muy a pecho	0	1
Abrir a alguien los ojos	0	1
Revolver el estómago	0	1
Tener hasta la madre	0	1
Dar lo mismo	0	1
No tener precio	0	1
Hablar por los codos	0	1
Valer madre/verga/verch/mierda	0	4
Pasarse (algo) de moda	0	1
Dar gusto	1	0
Cagar(la) [arruinarla]	2	0
cagar(se) en [para expresar desprecio]	3	8
Morirse de risa	0	1
*locuciones adjetivas		
A toda madre	1	0
Poca madre	1	0
Pan comido	0	1
De huevos	0	1
Del Diablo	0	2
De Dios	0	1
Tan de la chingada	0	1
Corta venas	1	0
Bien portado	1	0
Bien de la verga	2	0
Total	37	51

Tabla. Enunciados fraseológicos

	Hombres	Mujeres
¿Eso qué?	1	0
¿Qué pedo con…?	3	6
¿Qué onda con…?	1	0
Nada como	5	0
Sin palabras	1	0
No cualquiera	1	0
Sabes que el karma existe cuando	1	0
Casi lo mata	1	0
¿Qué más se puede pedir?	5	0
Vida eso que pasa	2	0
Salirse de control	1	0
Ya no hay niveles	1	0
Hacer (el día, la noche, la tarde, la mañana)	1	2
Para acabar	1	0
Mueve el mundo	1	0
¡Qué… ni qué nada!	0	1
¿Por qué demonios…?	0	1
Límite de mi tolerancia	0	1
¡Ya mátenme!	0	1
No porque me enamoro	0	2
Me quiero volver chango	0	1
¡Qué máquina ni qué 8/4!	0	1
Lloro de felicidad	0	1
La cereza del pastel	0	1
Ni el sol me calienta	0	1
Algo es algo	1	1
Pinche(puerca) vida	1	2
La vida no vale nada	1	1
Total	29	23

Tabla. Verbos

Verbos no psicológicos	**Hombres**	**Mujeres**
Acostumbrarse	1	0
Agradecer	1	2
Alucinar	0	1
Amenizar	1	0
Apedrear	0	1
Apestar	0	1
Aplaudir	1	0
Aportar	0	1
Arruinar	1	2
Ayudar	1	0
Cagar(arruinar)	2	0
Chingar(joder)	1	2
Chingar(quitar)	1	2
Compensar	1	0
Complicar	0	1
Confundir	0	1
Desmotivar	1	0
Discriminar	1	0
Destruir	0	1
Elogiar	1	0
Enfermar	0	1

Halagar	1	0
Ignorar	1	0
Impactar	0	1
Imponer	0	1
Incurrir	1	0
Inflar	1	0
Influir	1	0
Joder	1	1
Limpiar	1	0
Malacostumbrar	0	1
Maltratar	1	0
Matar	1	5
Masacrar	1	0
Merecer	5	0
Mejorar	0	1
Necesitar	1	0
Entender	1	1
Saber	0	1
Poder	0	1
Servir	1	2
Ñoñear	0	1
Purgar	1	0
Pecar	1	0

Presumir	1	1
Quejarse	2	0
Salvar	0	1
Reconciliar	0	1
Soñar	1	0
Superar	1	1
Ocupar	0	1
Solucionar	1	0
Regar	1	0
Recomendar	3	0
Rockear	2	0
Tumbar	1	0
Spolear	1	0
Tardar	1	0
Valer	1	2
Cagarse (de miedo)	1	0
Cansar(agotar)	0	1
Total	49	38

Verbos psicológicos	Hombres	Mujeres
Adorar	0	12
Amar	6	39

Detestar	0	3
Emputar	0	1
Encantar	2	4
Extrañar	8	16
Gustar	12	26
Hartar	0	1
Love	0	4
Aguantar	0	1
Molestar	0	2
Soportar	0	2
Odiar	4	20
Querer	11	22
Aburrir	2	4
Admirar	1	0
Agradar	2	0
Angustiarse	1	0
Antojar	1	0
Asustar	1	2
Avergonzarse	1	0
Cansarse (hartar)	7	3
Divertir	1	0
Doler	4	7
Envidiar	0	1

Enojar	1	1
Importar	1	6
Inquietarse	0	1
Sufrir	2	2
Sorprender	3	1
Deprimir	1	0
Enorgullecerse	1	0
Preferir	0	1
Desear	1	0
Espantar	1	0
Enamorarse	0	2
Emocionar	2	0
Care	1	0
Preocuparse	0	1
Disfrutar	4	2
Interesar	0	1
Inspirar	0	1
Total	83	191
Verbos psicológicos + verbos no psicológicos	132	229

Tabla. Sustantivos

	Hombres	Mujeres
Alegría	1	0
Aliento	1	0
Amor	1	7
Animal	0	1
Answer	1	0
Antojo	0	1
Asco	0	2
Apoyo	1	0
Artista	1	0
Arte	0	1
Audacia	1	0
Basura	0	2
Bazofia	3	0
Belleza	1	0
Beneficios	0	1
Bien	1	0
Rencor	0	1
Oro	1	0
Broma	1	0
Caca	0	1
Cagadero	1	0
Campeón(es)	1	1
Cáncer	1	0
Capacidad	0	1
Carajos	0	1
Catástrofe	1	0
Cinismo	1	0
Cielo	0	2
Clave	1	0
Confianza	1	0
Corrupción	3	0
Consuelo	0	1
Coraje	0	2
Culpa	0	2

Creatividad	1	0
Cristal	1	0
Criminalización	1	0
Daño	2	1
Decepción	0	1
Descaro	0	1
Desmanes	0	1
Dedicación	1	0
Deleite	1	0
Desastre	3	0
Desgracia	1	0
Desmadre	1	0
Despotismo	1	0
Dificultad	1	0
Dicha	0	1
Distorsión	0	1
Dios	0	1
Emoción	1	1
Enemiga	0	1
Erratas	1	0
Esfuerzo	1	0
Estrés	2	0
Espectáculo	0	1
Espectacularización	1	0
Esperanza	0	1
Esplendor	0	1
Extremo	3	0
Faltas	1	0
Fan	3	2
Felicidad	0	4
Fosa	1	0
Fuerza	1	0
Dignidad	1	0
Genio	0	2
Genialidad	1	0
Gloria	1	0
Gusto (-azo)	1	2

Gozo	0	1
Grosería	0	1
Hazaña	0	1
Héroe	1	0
Hijo de puta	0	1
Historia	2	0
Hit	1	0
Hope	1	0
Horror	1	0
Hueva	6	3
Idiotez	1	0
Ídolo	1	0
Incertidumbre	1	0
Incongruencia	1	0
Indiferencia	1	0
Infarto	1	0
Injusticia	0	1
Inmundicia	1	0
Interés	0	1
Impunidad	0	1
Impresión	0	1
Jedi	1	0
Jefe	1	0
Joya	1	0
Lástima	3	0
Límite	1	1
Locura	1	0
Maldad	0	1
Maldición	1	0
Mame (s)	2	1
Marranadas	1	0
Mamada(s)	0	3
Matadero	0	1
Masacre	1	1
Melancolía	0	1
Mérito	1	0
Miedo	2	7

Mierda	5	5
Miseria	1	0
Motivación	1	0
Milagro(s)	1	0
Nada	1	0
Narcisismo	1	0
Megalomanía	1	0
Motor	0	1
Muerte	0	1
Obsesión	1	0
Orgullo	1	0
Oso	0	3
Onda	0	1
Necedad	1	0
Paranoia	2	0
Pena	3	1
Pendejadas	2	7
Paciencia	1	0
Pánico	0	1
Pasión	0	1
Pedos	0	2
Pex	0	3
Plus	0	1
Pérdida	1	0
Pesadilla	1	0
Placer(es)	2	1
Pleasure	1	0
Poder	1	0
Poetambre	1	0
Porquería	0	1
Provocación	0	1
Problema(s)	6	3
Problems	1	1
Recomendación	1	0
Revoltijo	1	0
Risa	1	0
Represión	1	0

Riddikkulus	0	1
Rechazo	1	0
Sadness	1	0
Satisfacción	1	1
Sense	1	0
Sentido	2	0
Shame	1	0
Sueño	2	0
Suerte	0	1
Ternura	0	1
Tragedia	0	1
Trauma	0	1
Tranquilidad	0	1
Tristeza	0	1
Terror	1	0
Todo	1	0
Utopía	1	0
Vasallo	1	0
Vergüenza	1	1
Vicio	0	3
Violencia	1	0
Total	147	122

Tabla. Adjetivos

	Hombres	Mujeres
Absurdo	1	1
Aburrida	1	3
Adicto	2	0
Afortunada	0	1
Alfa	0	1
Alienado	1	0
Altiva	1	0
Aprobado	0	1
Ardiente	0	1
Arruinado	0	1
Asaltable	0	1
Asesino	1	0
Atractivo	0	1
Atroz	0	1
Avorazado	0	1
Bad	1	0
Bello(a)	4	6
Bendito	2	2
Best	0	2
Bien	0	1

Bobo	0	1
Bonito(a)	4	19
Bueno(a)	42	14
Bruta	1	0
Cabrón	3	2
Cagones	0	1
Cakil	1	0
Cansado	2	1
Caótico	0	1
Chaquetón	1	0
Chéveres	1	0
Chicles[9]	1	0
Chido	11	5
Chillones	0	1
Chingón	2	1
Chingados	0	1
Chulo	0	1
Claros	1	0
Clavas	1	0
Cobarde	1	1
Cómodo	0	2
Complicado	1	0

[9] Caso: "Tsa, esa morra está bien chicles"

Concurrida	0	1
Confundido	0	2
Contento	0	1
Conformista	1	0
Correcta	0	1
Corrupto	3	0
Comprometido	1	0
Crack	1	0
Creepers	1	0
Culposo	0	1
Cursi	0	4
Decente	0	1
Definitivo	1	0
Deliciosa	1	1
Deprimente	0	1
Deprimida	0	1
Demasiados	8	1
Descarado	1	0
Descuidada	0	1
Desgastados	0	1
Desgastante	1	0
Desmadroso	1	0
Desesperante	0	1

Diferente	0	1
Difícil	5	4
Divertido	2	3
Digna	0	1
Divina	0	1
Dulces	1	0
Duro	0	1
Efímero	2	0
Empoderada	1	0
Electorera	1	0
Enamorada	1	2
Encantadora	1	0
Enfermizas	0	1
Enojada	0	1
Ensimoso	1	0
Equis	0	1
Espantosa	1	0
Especial	1	2
Esperado	0	1
Espontaneous	0	1
Estresante	0	1
Estúpido(a)	3	4
Excelente	4	2

Fácil	4	6
Falaz	1	0
Falsa	1	0
Famoso	2	0
Fatua	1	0
Favorita	13	6
Favorite	0	1
Feliz	5	8
Feo(a)	3	5
Fino	1	0
Floja	0	1
Freaky	1	1
Firme	1	0
Fuerte	4	1
Genérico	1	0
Genial	1	0
Glorioso	0	1
Gracioso	2	1
Gran(de)	13	9
Grave	1	0
Guapo	3	7
Guilty	1	0
Guarro	1	0

Galán	0	1
Guay	0	1
Hermosa	5	12
Heroica	1	0
Hipócrita	0	2
Honesta	1	0
Horrible	4	2
Horrendo	1	0
Hostil	1	0
Idiota	2	3
Ignorante	2	0
Ignorado	0	1
Igual	0	1
Ideal	1	1
Happy	0	1
Huraña	0	1
Imbécil	1	1
Implícito(chingón, chido, padre)[10]	3	0
Impresionante	0	1
Importante	5	0

[10] El twittero utilizó expresiones superlativas pero omitió el adjetivo. Por ejemplo: “Eddie Palmieri es lo más” y “Esas empanadas eran lo más. Las extrañaré. :(”

Inalcanzable	0	1
Incapaz	0	1
Increíble	9	3
Indecisa	0	1
Indignado(a)	0	1
Infeliz	0	2
Injusto	0	1
Inmaduras	0	1
Inteligente	0	1
Interesada	0	1
Interminable	0	1
Intolerante	0	1
Introvertido	0	1
incómodo	2	0
Incompetente	2	0
Impuro	1	0
Inenarrable	1	0
Inepto	1	0
Informado	1	0
Inhumana	1	0
Inolvidable	1	0
Insulso	1	0
Interesante	1	0

Íntimo	1	0
Inútil	1	2
Invisible	0	1
Invaluable	1	0
Irresponsable	0	1
Jarcor	1	0
Jodida	0	1
Justo	1	0
Ladrón	2	0
Lamebotas	0	1
Lamentable	3	0
Leal	1	0
Lenta	1	0
Legal	0	1
Limpio	0	1
Lindo(a)	3	5
Listo	1	0
Loca	4	1
Machista	0	1
Madura	0	1
Malo(a)	10	6
Maestro	1	0
Mágico(a)	2	1

Mainstream	1	0
Maldito	8	9
Malévola	0	2
Mamila	0	1
Mamona	1	1
Maravilloso	0	2
Marica(ona,eta)	1	3
Mediocre	1	0
Mejor(es)	26	20
Mejorados	1	0
Memorable	1	1
Merecido	1	0
Menos	0	1
Molesto	0	2
Monótona	0	1
Morons	1	0
Muerto	1	0
Mugrosito	1	0
Mugriento	1	0
Naco	2	1
Necesario	0	2
Necio	1	1
Nefasto	1	0

Nopal	1	0
Ñoño(a)	1	3
Obsesionada	0	1
Obsesiva	1	0
Ojete	1	0
Ordinaria	1	0
Ofensivo	0	1
Oportuno	0	1
Orgulloso	0	5
Padre	2	0
Panes	1	0
Partidaria	0	1
Pedante	0	1
Patético	0	2
Perfecto	2	7
Pendejo	11	11
Peor	7	9
Perdedor	0	1
Perezoso	0	1
Perras	2	0
Pesado	3	0
Pésimo	1	2
Picado	0	1

Piltrafientas	1	0
Pinche(s)	15	13
Pobre	2	2
Popular	1	0
Posible	1	0
Preparada	0	1
Preocupado	1	0
Preocupante	1	0
Primordial	0	1
Problemático	0	1
Puto(s)	4	6
Querido	0	2
Radioactiva	1	0
Raro	3	2
Real	0	1
Remunerado	1	0
Rencoroso	0	1
Recomendable	0	1
Recomendado	1	0
Relajador	1	0
Rentable	1	0
Retrógrado	1	0
Rescatable	0	1

Resuelto	0	1
Revelador	0	1
Rico	1	5
Ridículo	5	4
Robusta	1	0
Romántico	0	2
Satisfied	0	1
Salvaje	1	0
Sano	2	0
Satírico	1	0
Seguro	1	0
Sencillo	1	0
Pseudohumana	0	1
Sola	0	1
Sensual	1	0
Sexy	1	3
Simple	2	2
Simulada	1	0
Sublime	0	1
Suficiente	2	1
Soquete	1	0
Suicidas	1	0
Talentoso	0	1

Tardado	0	1
Tedioso	0	1
Terrorífica	0	1
Terrible	2	0
Tierno	0	2
Tocado	0	1
Torpe	0	1
Tranquila	0	1
Trasher	0	1
Triste	7	8
Único	0	1
Verdadero	0	2
Vergas	1	1
Víbora	0	1
Vulgar	1	1
Tosco	1	0
Tramposo	1	0
Traumado	1	0
Vacío	2	0
Vendido	1	0
Válido	1	0
Troll	1	0
Weird	1	0

Tragón	1	0
Represor	0	0
Total	433	407

Tabla. Temas

	Hombres		Mujeres	
Personas	133	19.38%	210	29.16%
Música	66	9.62%	49	6.80%
Cine	46	6.70%	11	1.52%
Sí mismos	44	6.41%	57	7.91%
Cultura	4	0.58%	14	1.94%
Naturaleza	8	1.16%	10	1.38%
Lengua y literatura	36	5.24%	50	6.94%
Transportes	6	0.87%	2	0.27%
Escuela	22	3.20%	42	5.83%
Deportes	67	9.76%	21	2.91%
Política	48	6.99%	11	1.52%
Economía	6	0.87%	1	0.138%
Tiempo	26	3.79%	31	4.30%
Alimentos	8	1.16%	22	3.05%
Animales y plantas	3	0.43%	16	2.22%
Alcohol y drogas	10	1.45%	1	0.138%
Trabajo	8	1.16%	3	0.416%

Lugar	14	2.04%	12	1.66%
Entretenimiento	40	5.83%	27	3.75%
Acción	21	3.06%	42	5.83%
Astros y espacio	0		6	0.83%
Tecnología	13	1.89%	9	1.25%
Relaciones	3	0.43%	4	0.55%
Salud	2	0.29%	3	0.416%
Mente	6	0.87%	11	1.52%
Vida	32	4.66%	34	4.72%
Físico	2	0.29%	5	0.69%
Objetos varios	10	1.45%	7	0.97%
Personalidad	1	0.14%	4	0.55%
Fotografía	1	0.14%	3	0.41%
Tatuajes	0		2	0.27%
Total	686		720	

10.2 Apéndice II

Ji cuadrada de los elementos de opinión

Sustantivos contra el resto de elementos utilizados para opinar

	Hombres	**Mujeres**
Sustantivos	147	122
Resto de elementos opinativos	688	745
P-value = 0.05346		

H0= El uso de sustantivos no depende del sexo
H1= El uso de sustantivos depende del sexo
P-value mayor a 0.05, se acepta H0: el uso de sustantivos no depende del sexo.

Adverbios contra el resto de elementos utilizados para opinar

	Hombres	Mujeres		Hombres	Mujeres
Adverbios	21	12	Adv. + locuciones adv.	26	17
Resto de elementos opinativos	794	855	Resto de elementos	789	850
P-value = 0.1126				*P-value* = 0.1493	

H0= El uso de adverbios no depende del sexo
H1= El uso de adverbios depende del sexo
P-value mayor a 0.05, se acepta H0: el uso de adverbios no depende del sexo

Interjecciones contra el resto de elementos utilizados para opinar

	Hombres	Mujeres		Hombres	Mujeres
Interjecciones	16	23	Interj. +locuciones interj. + enunciados fraseológicos	57	56
Resto de elementos opinativos	799	844	Resto de elementos	758	811
P-value = 0.4371				*P-value* = 0.7335	

H0= El uso de interjecciones no depende del sexo
H1= El uso de interjecciones depende del sexo
P-value mayor a 0.05, se acepta H0: el uso de interjecciones no depende del sexo

Locuciones contra el resto de elementos utilizados para opinar

	Hombres	Mujeres
Locuciones	37	51
Resto de elementos opinativos	778	816
P-value = 0.2601		

H0= El uso de locuciones no depende del sexo
H1= El uso de locuciones depende del sexo
P-value mayor a 0.05, se acepta H0: el uso de locuciones no depende del sexo

Enunciados fraseológicos contra el resto de elementos utilizados para opinar

	Hombres	Mujeres
Enunciados fraseológicos	29	23
Resto de elementos opinativos	786	792
P-value = 0.481		

H0= Uso de enunciados fraseológicos no depende del sexo
H1= Uso de enunciados fraseológicos depende del sexo
P-value mayor a 0.05, se acepta H0: uso de enunciados fraseológicos no depende del sexo