
UNIVERSIDAD DEL BÍO-BÍO, CHILE

FACULTAD DE CIENCIAS EMPRESARIALES

DEPARTAMENTO DE SISTEMAS DE INFORMACIÓN


# RECURSO LÉXICO AFECTIVO PARA ESPAÑOL (REDPAL)


TESIS PRESENTADA POR ROBERTO VARELA MEDINA.


PARA OBTENER EL GRADO DE MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN


DIRECTORA: ALEJANDRA SEGURA NAVARRETE

CO-DIRECTORA: CLAUDIA MARTÍNEZ ARANEDA, UCSC.

## *Abstract*

*The task of analyzing subjectivity in texts is currently performed through approaches based on machine learning, lexicons and hybrid schemas. Lexicons used in the affection analysis only incorporate words which belong to each affective class, but they do not consider a measure to indicate the intensity of them. The purpose of this research is to design and build an affective lexical resource in Spanish based on an enriched lexicon. The resource represents the emotional intensity of every single word and this representation allows modifying and incorporating automatically in expansion processes. The obtained result is an affective tree that shows the intensity, facilitating the re labeling through crowdsourcing and subsequent regeneration and dynamic upgrades according to language evolution. This research is compared to a current affective lexicon in Spanish, evaluating its representation and impact on affective analysis performance. The results show the new resource improves the affective analysis and significantly reduces ambiguity in the affective class identification which expresses the analyzed phrase.*

## Resumen

Actualmente, la tarea de analizar la subjetividad en los textos es realizada a través de enfoques basados en *machine learning*, lexicones o esquemas híbridos. Los lexicones utilizados en el análisis de afectos solo incorporan palabras que pertenecen a cada clase afectiva y no contemplan una medida que indique la intensidad de ellas. Esta investigación tiene como propósito diseñar y construir un recurso léxico afectivo en español basado en un lexicón enriquecido, el cual represente la intensidad de la emoción de cada palabra y cuya representación permita modificar e incorporar de forma automática en procesos de expansión. El recurso obtenido es un árbol afectivo que muestra la intensidad, facilitando el re etiquetado a través de crowdsourcing y su posterior regeneración y actualización dinámica de acuerdo a la evolución del lenguaje. Esta investigación es comparada con un lexicón afectivo existente en español, evaluando su representación y su impacto en el desempeño del análisis afectivo. Los resultados indican que el recurso creado mejora el análisis afectivo y reduce significativamente la ambigüedad en la identificación de la clase afectiva que expresa la frase analizada.

# Índice

# Índice de Figuras

# Índice de tablas

# 1 Introducción

El análisis de subjetividad en textos es un área de gran interés hoy en día, basta con realizar una búsqueda sobre este tema en los sitios de publicación científica más conocidos y los resultados superan los miles en cada uno de estos. El análisis de subjetividad en texto, puede realizarse bajo dos enfoques, uno basado en *machine learning y*, otro en el uso de lexicones. Sin embargo, se ha analizado que los resultados mejoran cuando se complementan ambos enfoques [1] [2] [3]. Este trabajo se orienta en específico al análisis basado en lexicón en donde el factor crítico de éxito recae en el lexicón utilizado, es decir, en el tamaño, la calidad del etiquetado, la clasificación, el idioma y dominio. El idioma inglés cuenta con la mayor cantidad de recursos léxicos, el desarrollo de recursos o de lexicones en idiomas distintos al inglés como en [4] [5] ante la ausencia de un corpus extenso, un lexicón específico obtiene mejores desempeños que enfoques basados en aprendizaje, para el análisis de sentimientos. De los lexicones disponibles en idioma español, la mayoría se encuentran etiquetados sólo con polaridad [6]. También existen otros lexicones en inglés disponibles multilenguaje [7] [8], no obstante, no siempre es posible usar las traducciones literales, ya que éstas no necesariamente expresan lo mismo que las usadas en español, más aún si se consideran modismos o expresiones informales.

Los lexicones afectivos deben considerar que una misma palabra se utiliza para expresar más de una emoción, pero no necesariamente con la misma intensidad. Por ejemplo, la palabra *adoration*/adoración está clasificada en el lexicón *EmoLEX* [9] [10] [7] simultáneamente en las clases *fear*, *trust*, *anticipation* y *joy*. No obstante, en un lexicón etiquetado con intensidad, aunque la palabra sigue estando clasificada en las mismas clases, considerando la intensidad en una escala de 0-100, la palabra *adoration* expresa menor intensidad cuando se relaciona a las clases *fear* (10) y *anticipation* (28) en contraposición a las clases *trust* (28) y *joy* (46). La intensidad es una métrica que representa la relación de dos conceptos basados en sus relaciones jerárquicas [11].

Finalmente, los lexicones son representados, lógica y físicamente, en estructuras planas y archivos en formatos csv o txt. Lo anterior implica que no están representados considerando las relaciones semánticas entre las palabras y que tampoco facilitan su actualización en el tiempo. Las palabras de un lexicón construido desde hace una década, independiente de la forma como fue construido, no tienen la misma utilidad en el análisis de subjetividad. Tal como dijo Víctor García de la Concha, vigesimoctavo director de la Real

Academia Española, “La lengua es un ser vivo que nace, se desarrolla y, a veces, muere por desuso; como todo ser vivo”[1].

Para mejorar el procesamiento del lenguaje natural en español, es necesario disponer de un recurso léxico afectivo para esta lengua donde se incorporen las relaciones semánticas y la intensidad afectiva de las palabras en las clases emocionales. Además este recurso debe ser representado de forma que permita ser procesado computacionalmente tanto para incorporar y etiquetar nuevas palabras, así como debe ser capaz de complementarse con otros recursos disponibles como Multilingual Central Repository [12] u otros diccionarios online.

Esta investigación se enmarca en definir un recurso léxico afectivo a partir de un lexicón enriquecido propuesto en [13] etiquetado en las ocho clase afectivas de Plutchik [14]. Este recurso será construido comenzando con un proceso de expansión inicial a través de WordNet [15], para luego realizar los procesos de desambiguación, etiquetado de intensidad, normalización y generación del recurso léxico. Tal como fue mencionado antes, el recurso léxico RedPal estará diseñado para permitir reorganización, actualización y análisis que sea perdurable en el tiempo.

El informe se organiza de la siguiente manera: En el capítulo Objetivos e Hipótesis de Investigación se formulan la hipótesis, los objetivos generales, específicos y metodología de trabajo. En siguiente capítulo corresponde al marco teórico y trabajos relacionados, el capítulo contempla los conceptos definidos para realizar la revisión sistemática, aplicada al trabajo. A continuación, se exponen los principales conceptos de análisis de sentimientos, lexicones y análisis de emociones. Una vez definido el marco teórico, en el capítulo siguiente se describe el método de trabajo para recurso léxico. Con el recurso léxico ya construido, en el siguiente capítulo se describe el experimento de evaluación. Los últimos capítulos de informe contienen la discusión, trabajo futuro y conclusiones del estudio.

[1] http://udep.edu.pe/castellanoactual/sobre-la-creatividad-lexica-xenismos-y-calcos/

# 2 Objetivos e Hipótesis de Investigación

## 2.1 Hipótesis

La representación de un recurso léxico afectivo en español basado en la intensidad emotiva y procesable computacionalmente, permite mejorar el rendimiento del clasificador en el análisis afectivo.

## 2.2 Objetivos

### 2.2.1 Objetivo General

Definir un recurso léxico en español basado en la intensidad de la emoción expresada en cada palabra, que permita regenerar e incorporar nuevas palabras de forma automática, con el propósito de mejorar el análisis afectivo de textos.

### 2.2.2 Objetivos Específicos

- Analizar recursos léxico afectivo disponibles; información contenida, relaciones semánticas, estructura y formatos disponibles.
- Diseñar una estructura que soporte el recurso léxico que permita incorporar la intensidad de la emoción de cada palabra y considere criterios de transformación y regeneración para el enriquecimiento futuro del recurso.
- Validar la efectividad de la propuesta a partir de un experimento que utilizando el recurso léxico propuesto mejore el rendimiento del análisis afectivo, incorporando un proceso de retroalimentación que regenere e incorpore nuevas palabras de forma automática.
- Analizar objetivamente los resultados obtenidos y generar conclusiones y propuestas de trabajo futuro.

### 2.2.3 Alcances de la Investigación

El resultado de esta investigación sentará las bases para disponer de un recurso léxico integrado en idioma español para apoyar los procesos de análisis de subjetividad de la comunidad de investigadores en temas de análisis afectivo en textos y en específico del grupo de Investigación SoMoS. El recurso léxico solo contiene palabras con carga emotiva,

es decir no contiene todas las palabras del idioma. Dicho recurso estará disponible para su uso a través de una herramienta (o software) que implementa algoritmos de incorporación de nuevas palabras, re etiquetado, regeneración de la red.

### 2.2.4 Metodología de trabajo

Para alcanzar los objetivos de esta tesis se definieron las siguientes etapas:

- **Revisión bibliográfica**. Concluir la versión del estado del arte a partir de las preguntas de la investigación sobre análisis de emociones o afectivo en textos, lexicones disponibles (principalmente en español) y formas de representar el lexicón como recurso léxico. En las revisiones preliminares se ha vislumbrado que estos temas no están relacionados en el área de análisis de sentimientos, es decir, al realizar un análisis de sentimientos no se habla de la representación del lexicón, ni se definen estructuras básicas para mejorar el clasificador o representarlo como recurso léxico, por lo cual el protocolo de búsqueda se enfocará en más de un tema (análisis de sentimientos y representaciones del recurso léxico).
- **Análisis del lexicón base**. Corresponde al estudio de las relaciones semánticas entre palabras de cada clase afectiva representadas en *WordNet* [15]. Así también, se analiza la forma como la estructura o relaciones determinan la intensidad de las palabras.
- **Análisis del lexicón expandido en español.** Corresponde al análisis de las de las palabras del lexicón base. La forma como se genera el lexicón en español.
- **Análisis de las posibles representaciones.** Es necesario analizar las distintas representaciones lógicas posibles de la red léxica. Dicha red incorpora al menos las relaciones padre e hijo (IS-A), sinonimia, y de pertenencia a una clase o emoción.
- **Diseño de propuestas de representación del recurso léxico.** A partir del análisis previo se obtendrán distintas alternativas de representación tales como Grafo [16] [17] [18], Tesauro (ISO 2788) [17], diccionario *WordNet* [15] [19] [19].
- **Análisis crítico de las propuestas en base a criterios definidos**. Dichas representaciones deben ser evaluadas en base a criterios como la facilidad de incorporación de nuevas palabras con su clasificación e intensidad, soporte automatizado, actualización o regeneración del modelo, eficiencia en búsqueda, interoperabilidad con otros recursos léxicos en español.

- **Diseño del experimento y Evaluación de la propuesta.** La red léxica generada debe ser evaluada, para lo cual se diseñará un caso de estudio y se analizarán los resultados obtenidos para realizar los cambios o mejoras a la red.

# 3 Marco Teórico y Trabajos relacionados

En este capítulo se describe el marco teórico de esta investigación, el cual se construyó mediante una revisión sistemática de la literatura [20]. Se describirán los conceptos esenciales para la investigación como son el análisis de emociones, lexicones y por otra parte, se detallan los trabajos relacionados con las representaciones de estos recursos léxicos.

## 3.1 Marco conceptual

### 3.1.1 Análisis de sentimientos

A modo de introducción en el área de análisis de subjetividad es necesario aclarar algunos conceptos importantes [2] [21] [22] [1] [23] [24]:

- Sentimiento: una representación subjetiva de las emociones, privada del individuo que las experimenta; de manera similar a la emoción, tiene una duración a corto plazo.
- Emoción: respuestas discretas y consistentes a eventos internos o externos que tienen un significado particular para el organismo; la emoción tiene una duración a corto plazo.
- Estado de ánimo: un estado afectivo difuso que, en comparación con la emoción, suele ser menos intenso, pero de mayor duración.
- Afecto: un término abarcativo usado para describir los temas de emoción, sentimientos y estados de ánimo juntos.
- Opinión: Idea, juicio o concepto que una persona tiene o se forma acerca de algo o alguien.
- Subjetivo: Representa un sentimiento, no algo objetivo.
- Objetivo: Contiene información que se puede comprobar.
- Polaridad: Positivo, negativo, neutral.

En el área del análisis de sentimientos podemos encontrar las siguientes sub clasificaciones [2] [21] [22] [1] [23] [24]:

- Clasificación de la polaridad de la opinión: Determinar si el texto expresa un opinión positiva, negativa o neutral.

- Detección de subjetividad: Saber si un texto es subjetivo. Generalmente un texto subjetivo expresa una opinión personal.
- Detección de emoción: Tarea de detectar si un texto expresa una emoción o no. similar a detección de subjetividad.
- Clasificación de Emociones: Tarea de detectar en forma más precisa la o las emociones existentes en un texto, y clasificarlas en un subconjunto definido.

El análisis de sentimientos es una de las áreas que ha tenido un gran crecimiento en el área de las Ciencias de la Computación, basta con realizar una búsqueda sobre este tema en los sitios de publicación científica y los resultados superan los miles en cada uno de estos, en la Figura 3-1 podemos ver las publicaciones por año en una base de datos bibliográfica como es *Scopus*. La gran mayoría de estas publicaciones comienzan a aparecer a comienzos del siglo 20 [25], con la posibilidad de realizar análisis sobre la información proveniente la Web. Sin embargo las bases el análisis computacional de subjetividad en textos aparece en los 90's [25]. En la Figura 3-2 se puede ver la evolución de popularidad del término "*sentiment analysis*" del 2004 a la fecha en Google Trends[2], la información muestra popularidad del término.

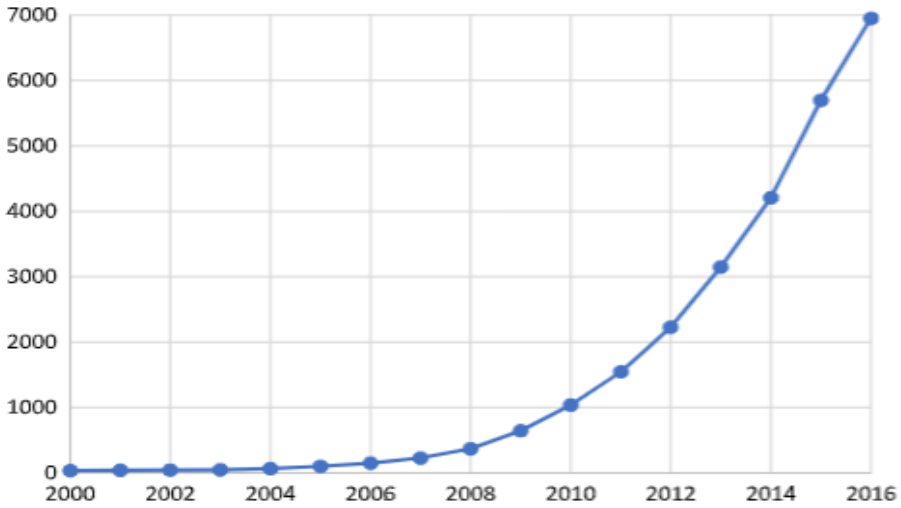



Figura 3-1 Numero de publicaciones en *Scopus* por año *[25]*.

En la Figura 3-1 se puede ver el constante incremento en las publicaciones relacionadas con el análisis de sentimientos.

[2] https://trends.google.com/trends/?geo=US

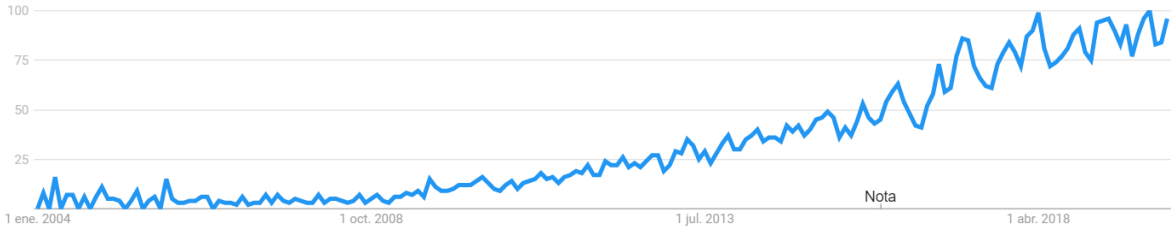



Figura 3-2 Popularidad del término “*sentiment analysis*”, en búsqueda de Google Trends.

Desde la Figura 3-2 y al igual que en la figura anterior se puede ver un constante incremento en el interés en las búsquedas relacionadas con el termino *sentiment analysis.*

Desde el punto de vista psicológico, las emociones humanas pueden ser identificadas y agrupadas según tipos de emociones, emoción, intensidad, y muchos otros parámetros. Los modelos de emoción son una forma estructurada para definir varias emociones humanas según algún puntaje, escala, rangos o dimensiones. Basado en diferentes teorías de las emociones, existen modelos que las dividen en dos clases: categórico y dimensional [23].

El modelo categórico define una lista de categorías y de emociones discretas. El modelo dimensional de emociones define dimensiones con algunos parámetros y emociones específicas de acuerdo con esas dimensiones. Se usan dos o tres dimensiones en la mayoría de los modelos dimensionales: "valencia" (indica la positividad o negatividad de una emoción), "excitación" (indica el nivel de estímulo de una emoción) y "dominación" (indica el nivel de control sobre una emoción) [23].

En cuanto a la clasificación de sentimientos existen muchos modelos que abordan dicha tarea, en [24] [26] se realiza un análisis exhaustivo del estado del arte en relación al análisis de emociones e introduce un modelo basado en las ocho emociones de Plutchik [14] teoría integradora basada en principios evolutivos. Las ocho emociones de Plutchik fueron seleccionadas por sobre otros modelos de seis emociones como Ekman, Shaver, Lovheim, Parrot ó Frijda. En la Figura 3-3 se muestran los modelos de emociones más usados según [24].

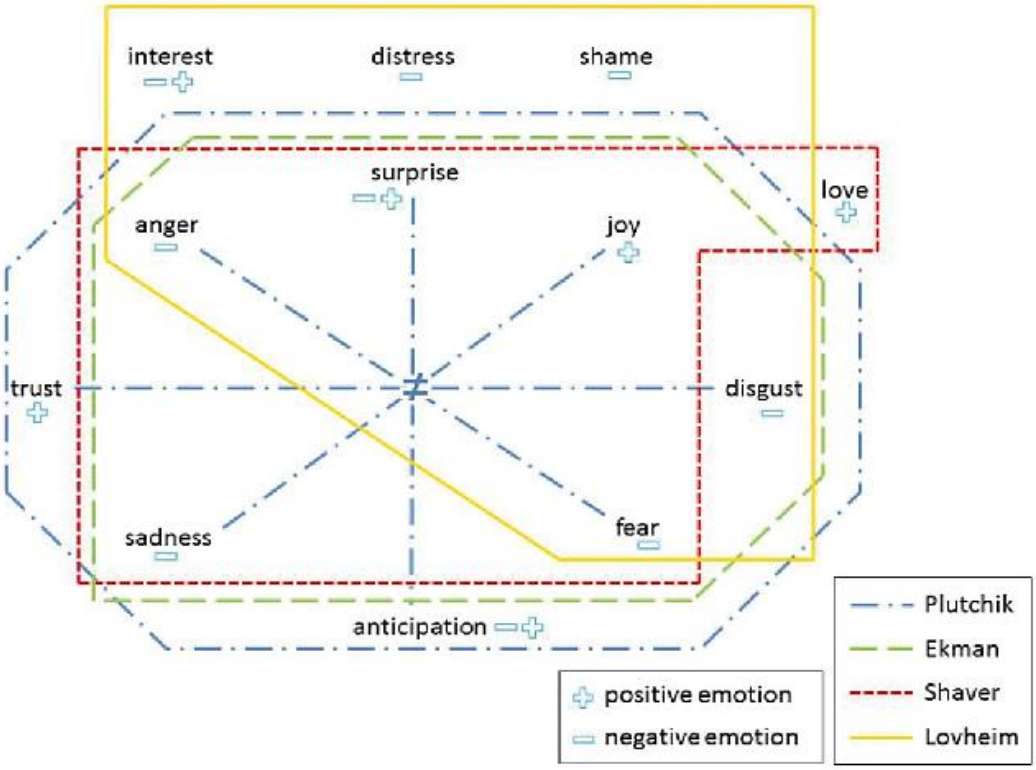



Figura 3-3 Los 4 modelos de emociones más frecuentemente usados [24].

Las 8 categorías primarias propuestas por Plutchik [14] son f*ear* (miedo), *surprise* (sorpresa), *sadness* (tristeza), *disgust* (aversión), a*nger* (ira), a*nticipation* (anticipación), *joy* (alegría) y *trust* (confianza) que corresponden a las emociones básicas. En el modelo dimensional de Plutchik [24], construido a partir de un análisis transcultural de la expresión emocional ante ciertos estímulos, el resultado su análisis fue ocho emociones primarias, en donde las emociones similares están a menor distancia del centro y posiciona las emociones en cuatro ejes opuestos: *joy* vs *sadness*, *anger* vs *fear*, *trust* vs *disgust* y *surprise* vs *anticipation*, además, define emociones positivas: *joy* y *trust*, emociones negativas: *anger*, *sadness*, *fear* y *disgust*, y emociones positivas y negativas: *anticipation* y *surprise*. El radio de la rueda indica intensidad, entre más cerca del centro mayor es la intensidad, la cual también se representa por el color, por lo tanto, mientras más intensa sea una emoción más se parecerá a la emoción básica y estará también más cerca del centro. Los espacios en blanco indican emociones complejas, formadas por las emociones básicas adyacentes. Otro aspecto a destacar es que cuanto mayor sea la intensidad de una emoción, el individuo estará más permeable a actuar en relación con ella ver Figura 3-4.

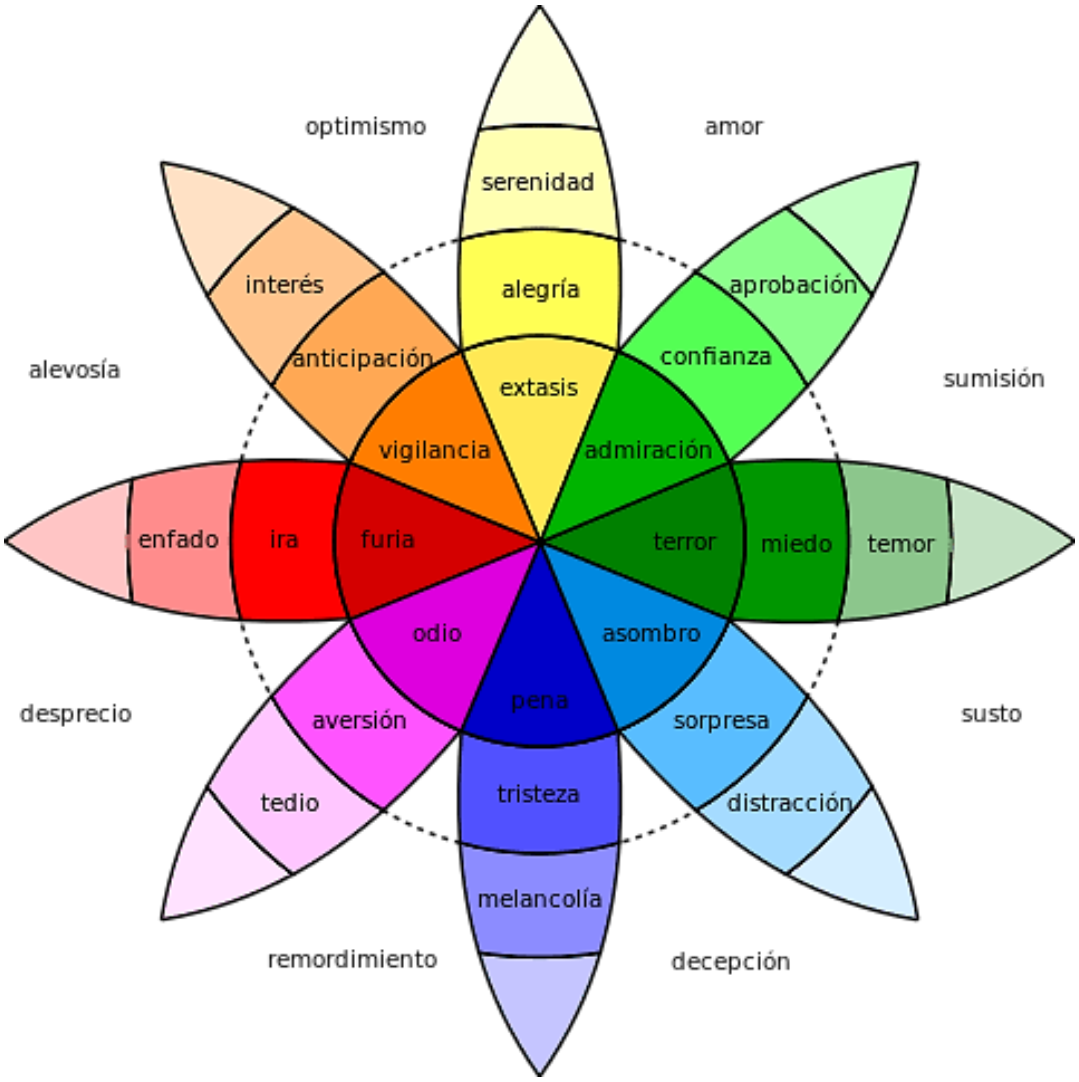



Figura 3-4 Rueda de emociones de Plutchik *[14]* traducida al español.

### 3.1.2 Lexicones

Durante la década de los 60's se originaron los lexicones, usados en un principio para caracterizar morfemas[3] de un idioma específico en una lista de palabras. A medida que crece la gramática transformacional se comienza a tratar como un componente del modelo del lenguaje que desempeña un papel auxiliar en la gramática [27] [28]. El lexicón es usado para apartar el conocimiento semántico en diferentes sistemas. La evaluación de jerarquías semánticas o lexicones siempre han sido un gran desafío [29].

Tal como se plantea en [30] un lexicón es el almacenamiento respectivo de la representación de una cantidad individual de conceptos con información sobre el mundo. Una palabra representa un cuerpo para un concepto o grupo de conceptos que aporta cierta información. Un lexicón es una estructura interna con diversas relaciones entre las unidades que existen en él. Un lexicón no es una estructura pasiva de almacenamiento del lenguaje, sino que un sistema dinámico que se organiza debido a la interacción continua entre el procesamiento y la estructuración de la experiencia.

Tal como se mencionó anteriormente un lexicón es una lista de morfemas en un lenguaje específico. Luego se incluyó un conjunto de reglas básicas que operan un lexicón. En el proceso de integración se pueden encontrar 2 tipos de unidades que definen cuan extensa es la generalización, estas unidades son los atributos diferenciados y componentes generalizados. Un lexicón varia en su grado de abstracción dependiendo de la cantidad de polisemia[4] que permita [27].

Un lexicón es un contenedor para almacenar una lista de temas o expresiones del lenguaje cuyo significado no es determinable a partir de los significados, por lo tanto, un usuario debe memorizar la combinación del lenguaje en forma y significado. Por otra parte, su organización dependerá de para qué fue diseñado [31].

Ejemplos de lexicón [31]:

- Un diccionario europeo organizado alfabéticamente.
- Un diccionario árabe organizado fonéticamente, comenzando con velares[5] y terminando con bilabiales[6].

[3] Unidad más pequeña de la lengua que tiene significado léxico o gramatical y no puede dividirse en unidades significativas menores.

[4] Fenómeno del lenguaje que consiste en que una misma palabra tiene varios significados.

[5] [Sonido consonántico, sonido vocálico] Que se pronuncia acercando la lengua al velo del paladar.

[6] [Sonido consonántico] Que se articula uniendo los labios.

- Una gramática que requiere una entrada léxica aleatoria según la clase morfológica que necesita acceder del lexicón a través de la categoría morfosintáctica[7] de la entrada del léxico.

Las estructuras en un lexicón jerárquico no solo son de gran importancia para la representación redundante de la información léxica, sino que también puede contribuir a la eficiencia del procesamiento del lenguaje natural [32] y a reducir la obsolescencia debido a su posibilidad de cambiar y adaptarse en el tiempo.

Los elementos del lexicón se pueden definir de acuerdo a distintas estrategias. Como por ejemplo siguiendo los enlaces (*IS_A),* padre hijo o es un, ordenando el conjunto de sus términos genéricos [33].

Según [28] un lexicón lingüístico contiene información sintáctica (función qué cumple la palabra en una frase) y semántica (significado), ejemplos ver Tabla 3-1.También evalúan la implementación del lexicón bajo orientación a objetos concluyendo que es muy útil para representar las palabras y sus propiedades, en un lexicón.

| Tipo de Información | En que Consiste |
|---|---|
| Sintáctica | Categoría<br>Rama<br>Genero<br>Formas irregulares<br>Conceptos de |
| Semántica | Relación inter-conceptos<br>Características esenciales<br>Ejemplos de estereotipos |

Tabla 3-1 Ejemplos de información sintáctica y semántica.

Existen distintos tipos de lexicón:

- Lexicones Gramáticos Figura 3-5.
- Lexicones de taxonomía de Conceptos comunes y específicos Figura 3-6.

Una jerarquía de herencia es usada para organizar un lexicón gramatical de gran tamaño y complejidad, resolviendo de manera eficiente la tarea [34]. Este está

[7] De la morfosintaxis o relacionado con esta disciplina lingüística.

implementado con OODB (base de datos orientada a objeto). La red generada permite representar la semántica de cada palabra de manera uniforme en nodos y enlaces.

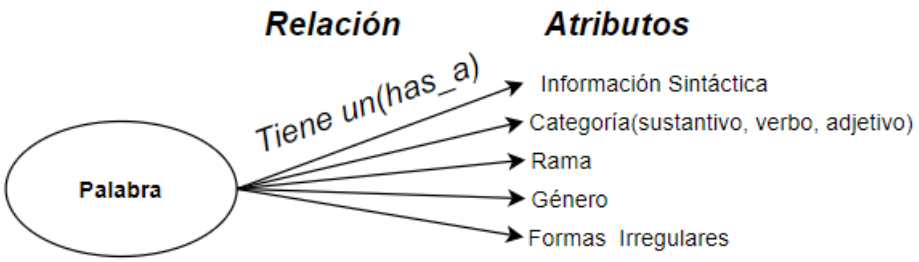



Figura 3-5 Características de un lexicón Gramático

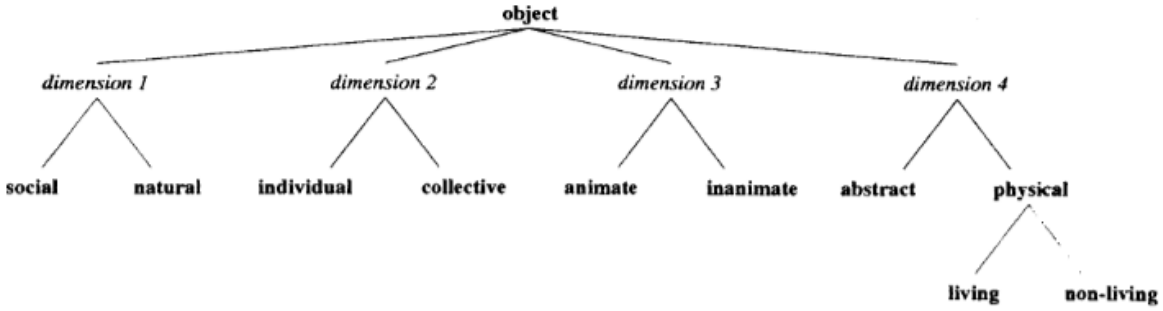



Figura 3-6 Ejemplo de extracto de una taxonomía para objetos(sustantivos) *[28]*.

En [32] se revisa una estructura jerárquica para soportar un lexicón, concluyendo que la estructura jerárquica analizada contribuye al desempeño, la reducción y resolución de ambigüedad del lenguaje natural. Finalmente, una estructura jerárquica ayudar en un proceso de análisis más rápido y eficiente reduciendo la redundancia en un lexicón y apoyo para la construcción de grandes lexicones.

En las teorías lingüísticas basadas en restricciones y proyectos lexicográficos, la información es organizada jerárquicamente, estas jerarquías son basadas en la "Valencia". Recursos léxicos como *WordNet*[8] *y FrameNet*[9], representan principalmente generalizaciones semánticas y sintácticas [35].

En [36] se presenta el diseño y acceso externo a un lexicón basado en la unificación de estructuras léxicas *HPSG-Style*(*Head-driven phrase structure grammar*, Gramática

[8] Base de datos léxica en inglés, que agrupa palabras en conjuntos de sinónimos llamados synsets. *https://wordnet.princeton.edu/*

[9] Base de datos léxica del inglés que es legible tanto por humanos como por máquinas, basada en ejemplos de anotaciones de cómo se usan las palabras en textos reales. *https://framenet.icsi.berkeley.edu/fndrupal/*

Sintagmática Nuclear) que no relacionan estructuras de superficie con estructuras profundas, y representadas en un Grafico acíclico dirigido (DAG por sus siglas en inglés *Direct acyclic graph*). La diferencia de usar el modelo relacional u Orientado a objeto para implementar el lexicón depende de la persistencia y objetivo buscado, donde el modelo orientado a objeto permitió representar de mejor forma un lexicón complejo ya que permitía representar como objetos los datos lingüísticos. Por otro lado, el modelo relacional no permitió almacenar el lexicón debido a las variables y recursividad de éste.

Un lexicón no es una estructura pasiva de almacenamiento del lenguaje, sino que un sistema dinámico que se organiza e interactúa continuamente entre el procesamiento y la estructuración de la experiencia. Un lexicón es una estructura interna con diversos links entre las unidades que existen en él [30].

## 3.2 Revisión sistemática de la literatura

Se realiza una revisión sistemática de la literatura con el propósito de conocer el estado del arte de los temas de investigación. Las temáticas a revisar son análisis de sentimientos, recurso léxico y sus representaciones y lexicones afectivos. A continuación, se detalla el protocolo para la revisión sistemática y sus resultados.

### 3.2.1 Preguntas de investigación

Se realizaron diversas preguntas enfocadas en los temas que aborda la hipótesis de esta investigación. Estas están orientadas al análisis de sentimientos, lexicones y las representaciones de estos y su impacto en recursos léxicos.

a) ¿Cuál es la relación de la estructura de representación lógica y física de un lexicón en el análisis de sentimientos?
b) ¿Cuáles son las clasificaciones afectivas, cuales son usadas en lexicones utilizados?
c) ¿Cuál es el impacto de la representación lógica y física de un recurso léxico en el análisis de sentimientos?

### 3.2.2 Protocolo de búsqueda

En la revisión de la literatura se usan nueve cadenas de búsqueda para responder a las preguntas de investigación.

Las cadenas traducidas usadas fueron: *sentiment analysis, lexicon, lexical resource, structure, Hierarchical.* Las combinaciones usadas se muestran en Tabla 3-2.

| *Combinación* | *Filtros de búsqueda* |
|---|---|
| *sentiment analysis - structure* | *Título del documento,*<br>*Resumen,*<br>*Palabras claves* |
| *sentiment analysis - lexicon* | |
| *sentiment analysis - Hierarchical* | |
| *sentiment analysis - lexical resource* | |
| *Lexicon - structure* | |
| *Lexicon - Hierarchical* | |
| *Lexicon - lexical resource* | |
| *lexical resource - structure* | |
| *lexical resource - Hierarchical* | |

Tabla 3-2 Combinaciones de las cadenas de búsqueda usadas.

Un ejemplo de búsqueda usada en el motor de búsqueda sería:

*[Abstract: "sentiment analysis"] AND [Abstract: structure] AND [Publication Title: "sentiment analysis"] AND [Publication Title: structure] AND [Full Text: "sentiment analysis"] AND [Full Text: structure]*

Aspectos de la búsqueda:

- En los buscadores de publicaciones científicas que dispongan de herramienta de búsqueda ingresar las combinaciones e incorporar los filtros de búsqueda dependiendo de las opciones de búsqueda avanzada que posea cada buscador. En caso del resultado de búsqueda supere las cien publicaciones, solo se revisarán las primeras cien que disponga el buscador.
- Los principales motores de búsqueda a usar en internet son: *Researchgate*, Web of *Science*, *Springer*, IEEE, ACM y *ScienceDirect.*
- Búsqueda cruzada: En los artículos encontrados buscar referencias bibliográficas que puedan ser de utilidad para la investigación.

### 3.2.3 Protocolo de Revisión

El protocolo de revisión de la literatura, posee las siguientes consideraciones:

(a) Normas de revisión: Recopilar las publicaciones completas impresas o su versión digital. Revisión exhaustiva de la introducción, resumen, tablas, imágenes, gráficos, formulas, conclusión y referencias.

(b) Criterios de inclusión: Se incluyen todas aquellas publicaciones que aborden los temas de análisis y respondan/aporten/ o estén en la línea de responder las preguntas de investigación.

(c) Criterios de exclusión: Se excluyen todas las publicaciones, que, a pesar de ser resultado de las combinaciones de búsqueda, no tienen información relevante o no abordan el tema de interés.

(d) Estrategia de extracción de los datos: Por cada estudio seleccionado, se realiza una lectura crítica con el objeto de extraer datos para el trabajo. Primero se lee la introducción, resumen, conclusión y referencias para saber:

- Introducción y Referencias: A qué comunidad contribuye.
- Resumen, Introducción y Conclusión: Sus aportes van en la línea de la investigación.
- Resumen, Conclusión e Introducción: Sus contribuciones son aportes.
- Cuerpo del Artículo: Incluir, separar y analizar la información útil en el estudio.
- Cuerpo del Artículo: Los experimentos, van en el marco de trabajo sobre el cual fueron desarrollados.
- Conclusión: Análisis crítico trabajos futuros.

(e) Estrategia de síntesis de los datos:

Los datos serán resumidos de acuerdo a los siguientes temas: Lexicón y estructura de lexicón, Detección de emociones, Análisis de sentimientos.

### 3.2.4 Selección de Estudios Primarios

Se realiza la búsqueda en los sitios *Researchgate*, Web of *Science*, *Springer*, IEEE, ACM y *ScienceDirect,* en donde se recopilaron 44 artículos en inglés, adicionalmente se recopilaron 29 artículos de distintas fuentes para abordar de mejor forma los distintos temas tratados en esta investigación.

### 3.2.5 Selección de Estudios Secundarios

De los 73 estudios seleccionados en la etapa primaria, se procedió a leer en detalle cada publicación de acuerdo al protocolo de revisión descrito. Además, se desecharon los trabajos que no indicaran claramente la forma o estructura que organizaban los lexicones. Finalmente se seleccionan 58 con los que se conforma el marco teórico y trabajos relacionados. En la Figura 3-7 se puede ver un resumen de las publicaciones incluidas y descartadas por año y en Figura 3-8 se puede ver un resumen de las publicaciones seleccionadas por temas de interés.

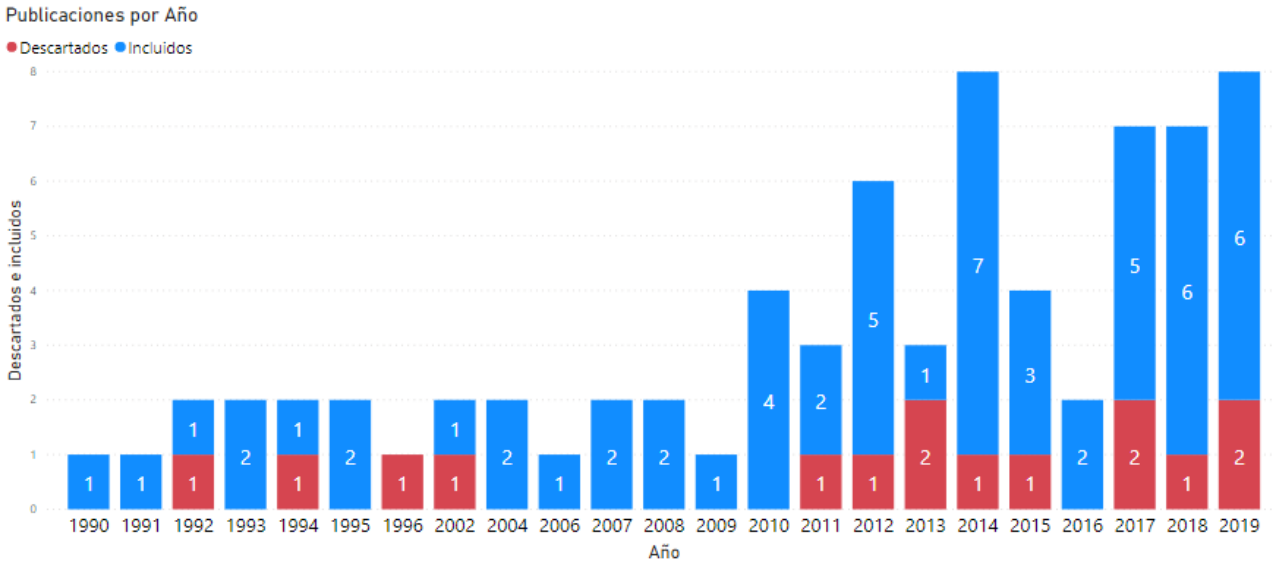



Figura 3-7 Resumen de publicaciones por año.

En donde 31 de los 58 artículos seleccionados se encuentra en los últimos 5 años (2014-2019).

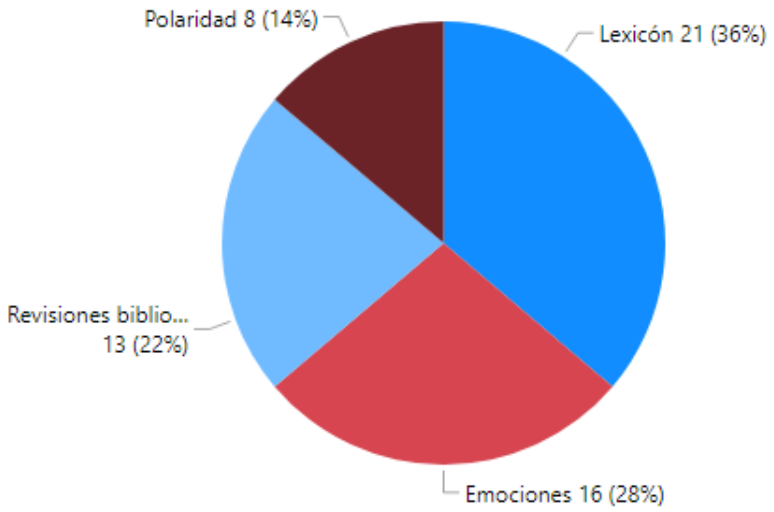



Figura 3-8 Resumen por temas de interés desde las publicaciones seleccionadas.

Como se observa en la Figura 3-8, el mayor porcentaje de artículos incluidos corresponde al tema Lexicón con 36%.

## 3.3 Marco Teórico

### 3.3.1 Análisis de emociones basado en Lexicón

Del análisis de la literatura [21] [22] [23] [24] podemos destacar que el “análisis de sentimientos” ha sido tema de estudio desde mediados de los 90. Sin embargo, no existe una unificación de las tareas o términos en esta área de estudio. Por ejemplo, el “análisis de sentimientos”, “clasificación de sentimientos” y la “clasificación de polaridad", se utilizan para abordar el mismo concepto “análisis de sentimientos”, asociado a determinar si un texto expresa una opinión positiva, negativa o neutral, también llamado clasificación de polaridad. Por otra parte la “clasificación de emociones” o “detección de emociones” o “minería de opinión” se enfoca en detectar la emoción o emociones existentes en un texto y clasificarlas generalmente en un modelo de emociones [24].

Según [1] las áreas de crecimiento en la detección de emociones son:

- Trasferencia de Aprendizaje.
- Detección de Emoción.
- Construcción de Recursos.

- Crecimiento en lenguajes distintos al inglés.

Los lexicones son muy útiles porque brindan información previa sobre el tipo y la fuerza de la emoción que conlleva cada palabra o frase. Los trabajos de minería de opinión usan poco los lexicones. [24].

En [23] [1] se realiza un análisis a las principales características estudiadas en clasificaciones de emociones, este análisis se realiza sobre 28 publicaciones, ver Tabla 3-3.

| Características | Publicaciones | Descripción |
|---|---|---|
| *Keyword Based* | 4 | Enfoque intuitivo y directo. La idea es encontrar los patrones similares a palabras clave de la emoción y hacerlas coincidir. |
| *Lexicon Based* | 7 | Es similar al enfoque *Keyword Based* solo usa una lista de palabras contenidas en el lexicón. |
| *Supervised-Learning-based* | 23 | A través de *Machine Learning supervisado o no supervisado,* diseñando modelos para entrenar un clasificador *Naive Bayes, Support Vector Machine, Decision Tree*, entre otros modelos. |
| *Semantic-based method* | 7 | Evalúa la similaridad semántica de las palabras. |
| *Hybrid method* | 6 | Combina los enfoques *Keyword Based, Lexicon Based y Supervised-Learning-based.* |
| *Generation of new lexicon* | 1 | Creación de un nuevo lexicón. |
| *Works with English language* | 23 | Trabajos en inglés. |
| *Works with other languages* | 6 | Trabajos en idioma distintos al inglés. |
| *Detects single emotion* | 28 | Encontrar la emoción predominante. |
| *Detects multiple emotions* | 0 | Encontrar más de una emoción. |
| *Emotion intensity detection* | 1 | Análisis por polaridad |
| *Generation of new data corpus* | 4 | A través de una metodología para etiquetar emociones y luego clasificarlas. |
| *Word-level detection* | 23 | Análisis por palabra. |
| *Sentence-level detection* | 6 | Análisis por oración buscando la orientación semántica de esta. |
| *Document-level detection* | 1 | Análisis de la emoción predominante en él documento. |

*Tabla 3-3 Resumen características estudiadas en la clasificación de emociones [23] [1].*

A partir del análisis de la literatura se construyó un resumen con las publicaciones que respondían las preguntas de investigación, referentes a análisis de sentimientos por polaridad y lexicón, el resumen de las 8 publicaciones encontradas en el tema se puede ver en la Tabla 3-4.

| Método Creación | Idioma | Tamaño Palabras | Clasificación | Escala o Valencia | Formato-Representación | Formato-archivo |
|---|---|---|---|---|---|---|
| Usando extracción automática desde un diccionario de subjetividad, realizan triangulación y expansión de un lexicón [37]. | inglés-español | *ND* | Positivo, negativo, altamente negativa y altamente negativa. | ND | *ND* | *ND* |
| Método automático desde tweets que incluyan el #hashtags del sentimiento. Separando en contextos negativos y positivos [38]. | inglés | 45000 unigramas positivos -9000 unigramas negativos | positiva, negativa | más negativo al más positivo | *<term> <tab> <score> <tab> <Npos> <tab> <Nneg>* | txt |
| Métodos automáticos y semiautomático basado en métodos de Rocchio y algoritmos SVM y revisión nativa [39]. | inglés | 6.000 | Positiva, negativa. | 0-1 | # POSID, PosScore{0,1}, NegScore{0,1} | XML synset y XML intensidad |
| Anotación automática de todos los *synsets* de *WordNet* [40] [41]. | inglés | 117.000 | positiva, negativa, neutral. | 0-1 | *<ID> <PosScore> <NegScore> <SynsetTerm> <Gloss>* | txt |
| Generan un lexicón a partir de la iteración de otros ya existentes [5]. | árabe | 4691 | positiva, negativa, neutral. | positivo-negativo | <palabra><positivo-negativo> | txt |
| A partir de un corpus etiquetado generan un lexicón de n-gramas [42]. | inglés | *ND* | positivo y negativo | 4 al -4 | *ND* | *ND* |
| A partir del navegación del árbol de *Wordnet*, construyen un grafo léxico afectivo, propagando afectivamente del árbol y con los pesos de las palabras calculan el valor positivo y negativo [16]. | inglés | *ND* | positivo y negativo | ND | *ND* | grafo y matrices |

*ND: información no disponible.*

*Tabla 3-4 Resumen de análisis de sentimientos por polaridad.*

Como conclusión del análisis detallado en la Tabla 3-4 se puede destacar que :

- 75% de los análisis realizados corresponde al idioma inglés, ver Figura 3-9.
- Existe una multiplicidad de formatos de representación de archivo, que contienen la representación del lexicón.
- Más del 40% de los trabajos los lexicones resultantes están en formato de archivo texto plano (*txt*), ver Figura 3-10.

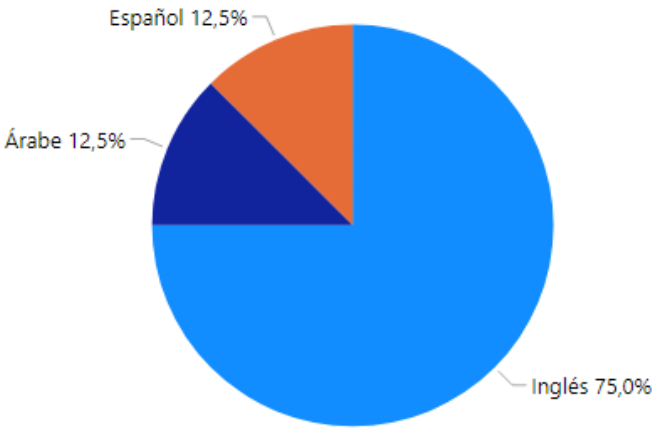



Figura 3-9 Lexicones por Idioma en trabajos analizados en la Tabla 3-4

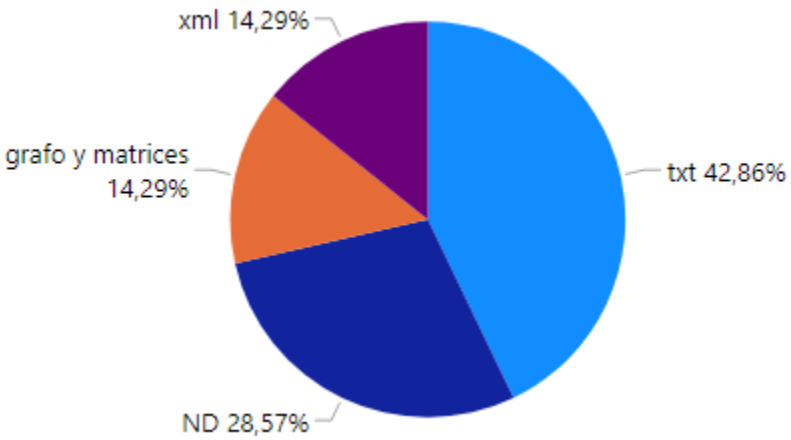



Figura 3-10 Lexicones por formato de archivo en los trabajos Tabla 3-4.

A partir del análisis de la literatura se construyó un análisis con las publicaciones que respondían las preguntas de investigación, referentes a análisis de clasificación de emociones o detección de emociones, el resumen de las 15 publicaciones encontradas en el tema se puede ver en las tabla Tabla 3-5.

| Método Creación | Idioma | Tamaño Palabras | Clasificación | Escala o Valencia | Formato Representación | Nota | archivo |
|---|---|---|---|---|---|---|---|
| Extraen las emociones automáticamente desde los #hashtags de *twitter* generando unigramas, asociando con las emociones [43]. | Inglés | 16.862 | *Plutchik(anger, anticipation, disgust, fear, joy, sadness, surprise, trust) más positivo y negativo)* | 0-1 | <unigram> <tab><AffectCategory> | corpus | txt |
| Anotaciones manuales por crowdsourcing en M*echanical Turk*. *Best–Worst Scaling*(BWS) Mejor-peor escalamiento [10] [44] [7]. | Inglés | 141.821 | *Plutchik(anger, anticipation, disgust, fear, joy, sadness, surprise, trust) más positivo y negativo)* | 0-1 | <term> <NearSynonyms> <tab> <AffectCategory> <tab><AssociationFlag> | Cada palabra esta 10 veces 1 vez por cada emoción + una positiva y otra negativa | txt |
| Anotación afectiva automática por crowdsourcing de una red social de noticias [8]. | Inglés | 37.000 | *Afraid, amused, angry, annoyed, dont care, happy, inspired, sad* | 0-1 | matriz: palabra x score por emociones | | múltiples txt |
| Anotaciones manuales crowdsourcing. *Best–Worst Scaling*(BWS) Mejor-peor escalamiento [45] [46]. | Inglés | 20.000 | *Valence{positive/pleasure, negative/unpleasure}; arousal {active/excited,sluggish/calm} dominance{powefull/strong; poweless/weak}.* | 0-1 | <Word> <Valence> <Arousal> <Dominance> | | txt |
| Anotación automática afectiva [47]. | Inglés | 5.000 | 6 emociones {*anger, fear, joy, sadness, surprise, disgust*} | 0-1 | <Concepts><Anger> <Disgust><Joy> <Sad><Surprise><Fear> | | xls |
| *ND: información no disponible.* | | | | | | | |

*Tabla 3-5 Resumen de análisis por clasificación de emociones.*

| Método Creación | Idioma | Tamaño Palabras | Clasificación | Escala o Valencia | Formato Representación | Nota | archivo |
|---|---|---|---|---|---|---|---|
| Desde diversos orígenes lingüísticos crean un vocabulario que modela las emociones y el proceso de análisis en una ontología multilenguaje y múltiples tipo análisis de emociones y polaridad [48]. | multilenguaje | *ND* | Distintos modelos: Fridja, Plutchik, Big6, entre otros. | | *ND* | | *ND* |
| Extraen las emociones automáticamente desde los #hashtags de *twitter* generando *unigramas*, asociando con las emociones [49]. | Inglés | 16.862 | *anger, anticipation, disgust, fear, joy, sadness, surprise, trust* | 0-1 | <unigram> <tab><AffectCategory> | corpus | txt |
| Anotaciones manuales. *Best-Worst Scaling* [50]. | Inglés | *ND* | *4 emotions {Anger, fear, joy, and sadness}* | 0-1 | <Term><score> <affectdimension> | | txt |
| Primero selecciona una rama afectiva desde *WordNet y luego se extiende a través de anotación manual* [51]. | Inglés | 4.000 | *anger, fear, joy, sadness, surprise, disgust* | 0-1 | <Term><affective-Weight> | (semantic affinity) | xml |
| Anotaciones manuales por crowdsourcing en Mechanical Turk. Mejor-peor escalamiento [9]. | Inglés | 1000+ | *anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, positivo y negativo* | 0-1 | <term>-- <NearSynonyms><tab> <AffectCategory><tab><AssociationFlag> | Encuestas de prueba para el crowdsoursing | txt |
| *ND: información no disponible.* | | | | | | | |

*Tabla 3-5 Resumen de análisis por clasificación de emociones (continua).*

| Método Creación | Idioma | Tamaño Palabras | Clasificación | Escala o Valencia | Formato Representación | Nota | archivo |
|---|---|---|---|---|---|---|---|
| Anotación afectiva automática por crowdsourcing de una red social de noticias [52]. | Inglés | 37.000 | *Afraid, amused, angry, annoyed, dont care, happy, inspired, sad* | 0-1 | matriz: palabra x score por emociones | | múltiples txt |
| Encuestas a estudiantes de sicología [6]. | Inglés español | 1.034 | Medida pictográfica SAM | Valencia(agradable-desagradable) Excitación(excitado-calmo) Dominio (fuera de control-control) | <numero> <palabra_ingles> <palabra_español> <evaluacion_afectiva> <indices_sicolinguisticos> | | data base |
| *ND: información no disponible.* | | | | | | | |

*Tabla 3-5 Resumen de análisis por clasificación de emociones (continua).*

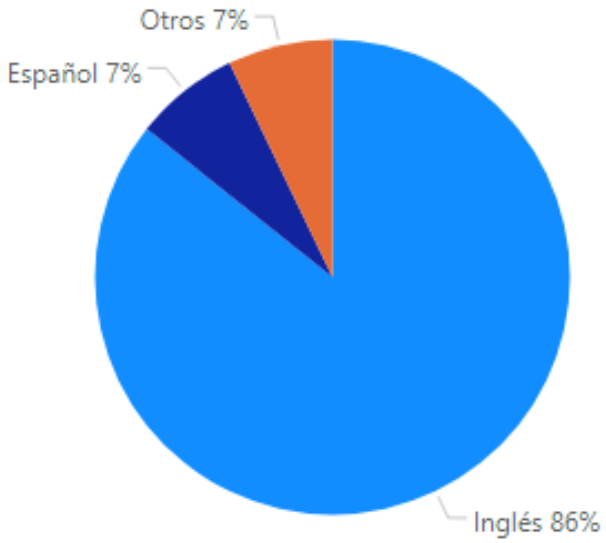



Figura 3-11 Clasificación de Emociones por idioma desde Tabla 3-5.

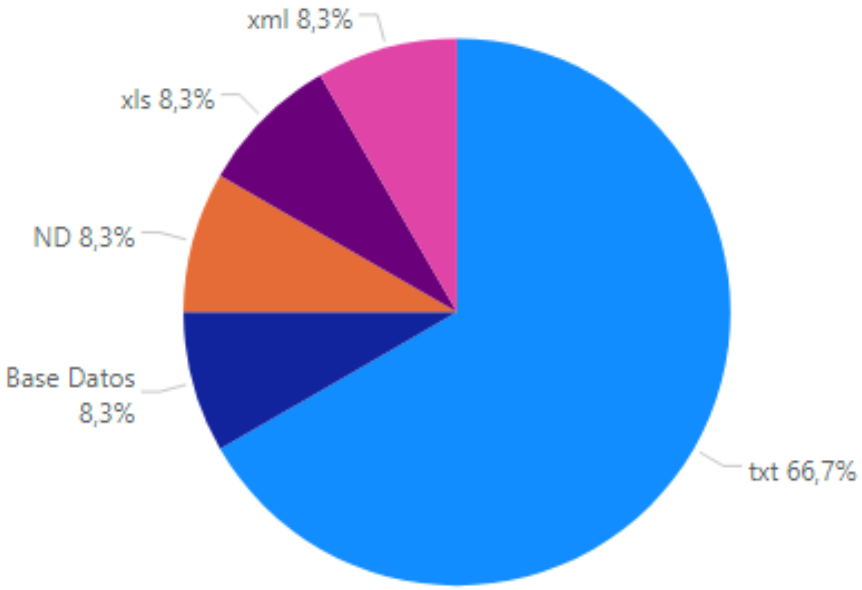



Figura 3-12 Clasificación de Emociones por formato de representación archivo desde Tabla 3-5.

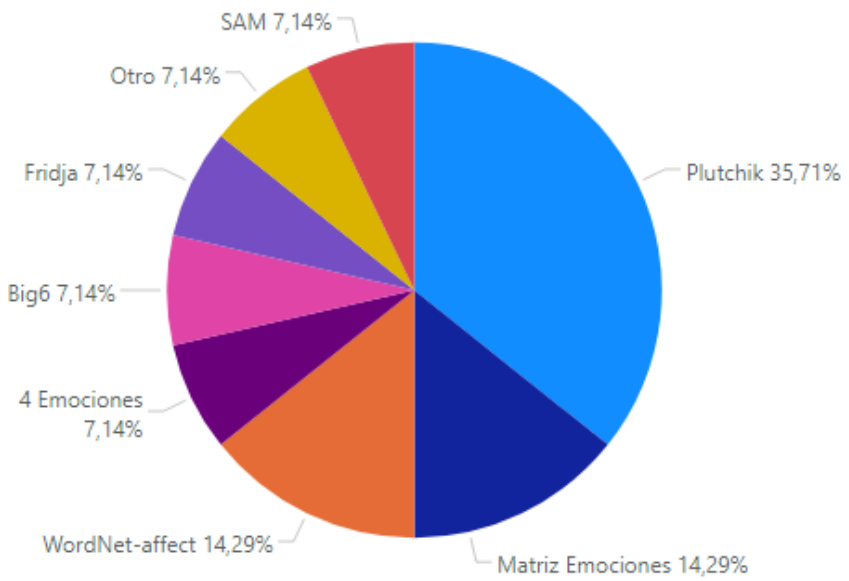



Figura 3-13 Clasificación de emociones por modelo desde Tabla 3-5.

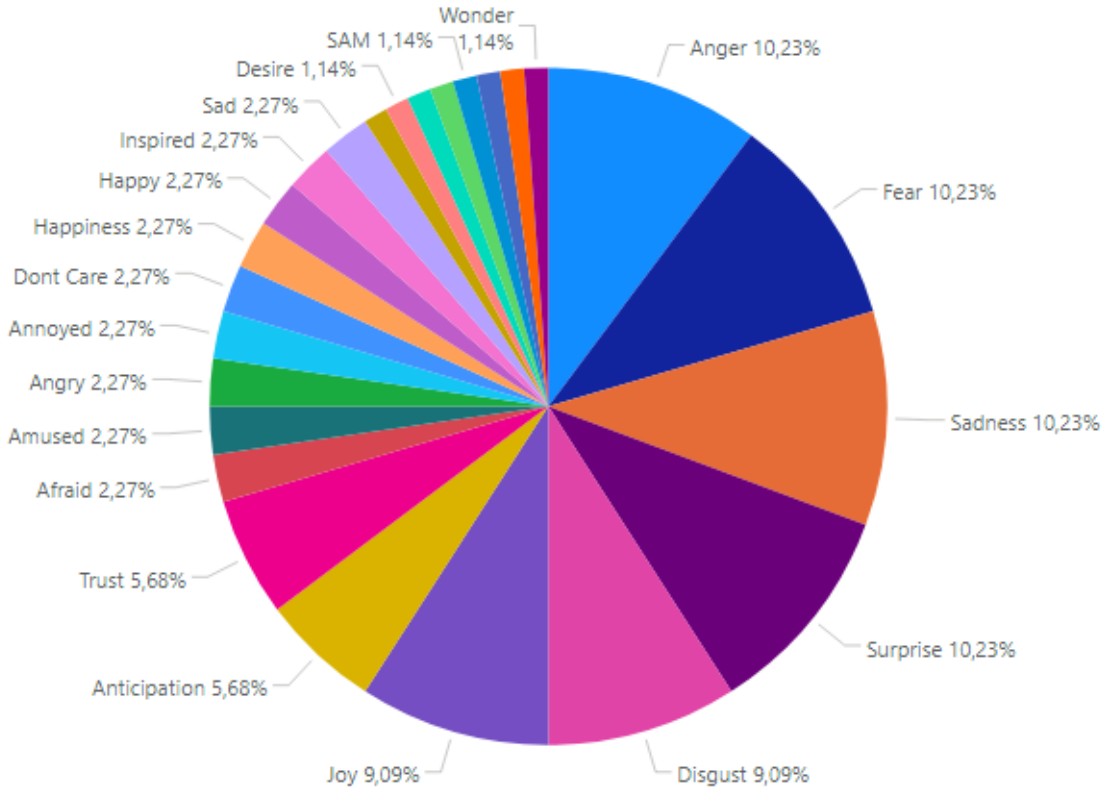



Figura 3-14 Clasificación de emociones por emoción desde Tabla 3-5.

Como conclusión de este análisis podemos decir:

- El 86% de los análisis realizados corresponde al idioma es inglés, ver Figura 3-11.

- Aunque existe multiplicidad de formatos de representación del archivo, el formato *txt* es el más usado con un 66%, ver Figura 3-12.
- Existe una gran cantidad de emociones usadas para realizar el análisis, pero son las pertenecientes del modelo de Plutchik son usadas con mayor frecuencia para realizar análisis (35%), ver Figura 3-13 y Figura 3-14.

### 3.3.2 Representaciones de los recursos léxicos en la literatura

Un lexicón de sentimientos es un recurso léxico que contiene información acerca de las implicancias emocionales de las palabras. En el estudio [39] se usa un grafo semántico de *WordNet* y *SentiWordNet*: recurso construido a partir de la clasificación automática de los *synset* de *WordNet* y recalcula la polaridad (positiva, negativa, objetiva).

Unos de los recursos léxicos en inglés más completos hoy es *WordNet* [15] un diccionario que incluye los sustantivos, verbos y adjetivos del inglés que se organizan en conjuntos de sinónimos(*synset*), cada uno de los cuales representa un concepto léxico subyacente. Las diferentes relaciones vinculan conjuntos de sinónimos. En la estructura de *WordNet* [15] la principal relación entre las palabras son los sinónimos, aunque también existen relaciones subordinadas a partir de tres relaciones semánticas: hiponimia[10], meronimia[11] y antonimia[12]. Al incluir todos los conceptos se obtiene una red interconectada o jerarquía bajo la relación (*IS_A)* (es un) y (*HAS_A*) (tiene un), la que ha sido suficientemente flexible para crecer y cambiar a través del aprendizaje.

Otro trabajo que usa las jerarquías para representación de un lexicón es [29], donde realiza una agrupación de hiperónimos[13] e hipónimos[14]para formar clases, lo que permitió una base léxica desambiguada, luego para cada clase se realiza una búsqueda de los hiperónimos e hipónimos que existan en otras clases con objeto de construir la jerarquía, y por último realizan una expansión a través de patrones en el texto.

En [48] se define una ontología para representar un vocabulario de emociones y un proceso de análisis de las mismas, en inglés. Para el análisis de emociones define un set de emociones en las que se incluyen las 8 emociones de *Plutchick* [14].

[10] Relación entre un hiponimo y otra palabra en cuyo significadao se encuetra englobado el del hiponimo.
[11] Palabra cuyo significado constituye una parte del significado total de otra palabra.
[12] Relación de oposición entre los significados de dos palabras.
[13] Palabra cuyo significado engloba el de otra u otras.
[14] Palabra cuyo significado esta englobado en el de otra.

La propuesta [16] construye un grafo de puntaje léxico basado en estructura *WordNet* (*IS_A)* de hiponimia y (*HAS_A)* de hiperonimia. En la Figura 3-15 se muestra un ejemplo de la jerarquía de hiponimia e hiperonimia.

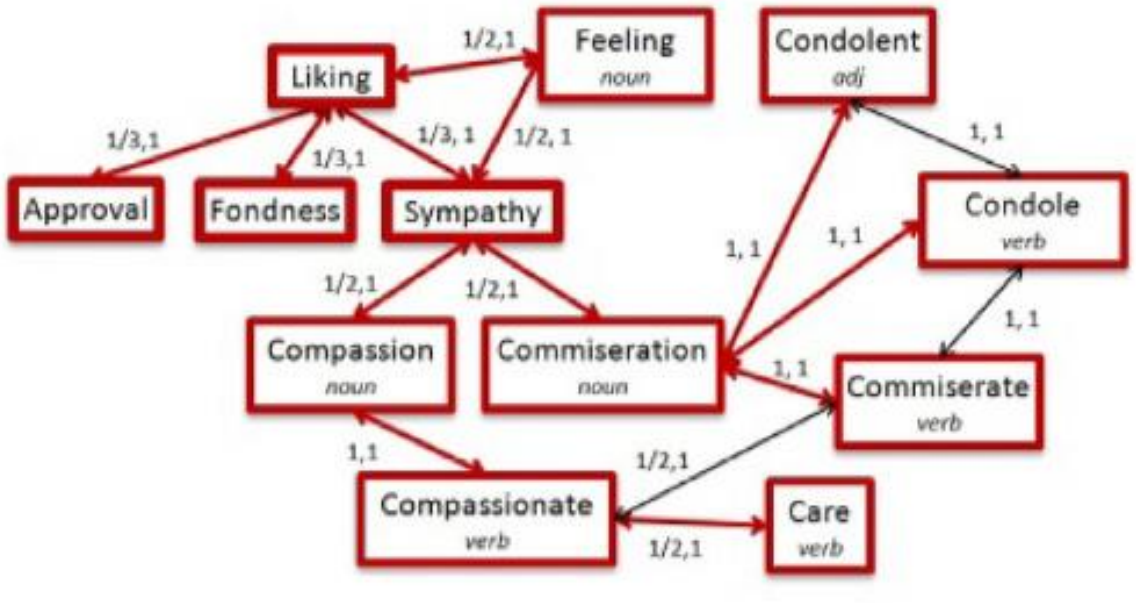



Figura 3-15 Ejemplo de la jerarquía de hiponimia e hiperonimia de *[16]*.

## 3.4 Conclusión del capítulo

Al realizar la revisión sistemática de la literatura se pudieron identificar los principales conceptos relacionados con lexicón y sus formas de organización, y *sentiment analysis*.

Se observa que:

- Las bases y la guía para el uso de lexicones son bastante antiguas.
- La forma de organización y las estructuras de un lexicón, está determinada por el uso que tendrá un lexicón.
- Un lexicón que contenga información sintáctica, semántica, reglas, generalizaciones y restricciones, necesitan una estructura que permita representar las palabras y propiedades, que serán mejor representadas en relaciones y/o jerarquías. Donde una implementación orientada a objeto o un modelo relacional son las mejores opciones y la elección de uno u otro dependerá de la persistencia buscada.

- Sobre las representaciones se pudo comprobar que los recursos léxicos más completos manejan distintos tipos de relaciones entre sus elementos lo que permite ser flexibles y, cambiar o expandirse.
- En *sentiment analysis,* existe un amplio interés y con mucho desarrollo tanto en áreas, tareas y características.
- La disponibilidad de un lexicón emocional específico y adecuado mejorar el reconocimiento de emociones [13].
- Los lexicones solo traducidos no mejoran el reconocimiento de emociones, es así como él foco debe estar en la construcción de lexicones para las necesidades particulares y abarquen las diferencias culturales con una énfasis en crear recursos en idiomas distintos al inglés [53].
- Los recursos en su gran mayoría están en formato de texto plano(*txt*) el cual, o no está disponible para uso o sólo está disponible para descarga, no existiendo un recurso el cual se pueda expandir y/o corregir con algún tipo de aporte científico o plataforma/repositorio que apoye estas tareas.
- El modelo de emociones de *Plutchik* es el más usado para el análisis de sentimiento y a su vez las emociones asociadas a este modelo también son las más usadas ya que también están incluidas en otros modelos de clasificación de emociones.

# 4 Método de trabajo para la creación del recurso léxico

En este capítulo se describe la forma en que se construyó el lexicón afectivo expandido. Este proceso puede ser representado a través del siguiente esquema, ver Figura 4-1.

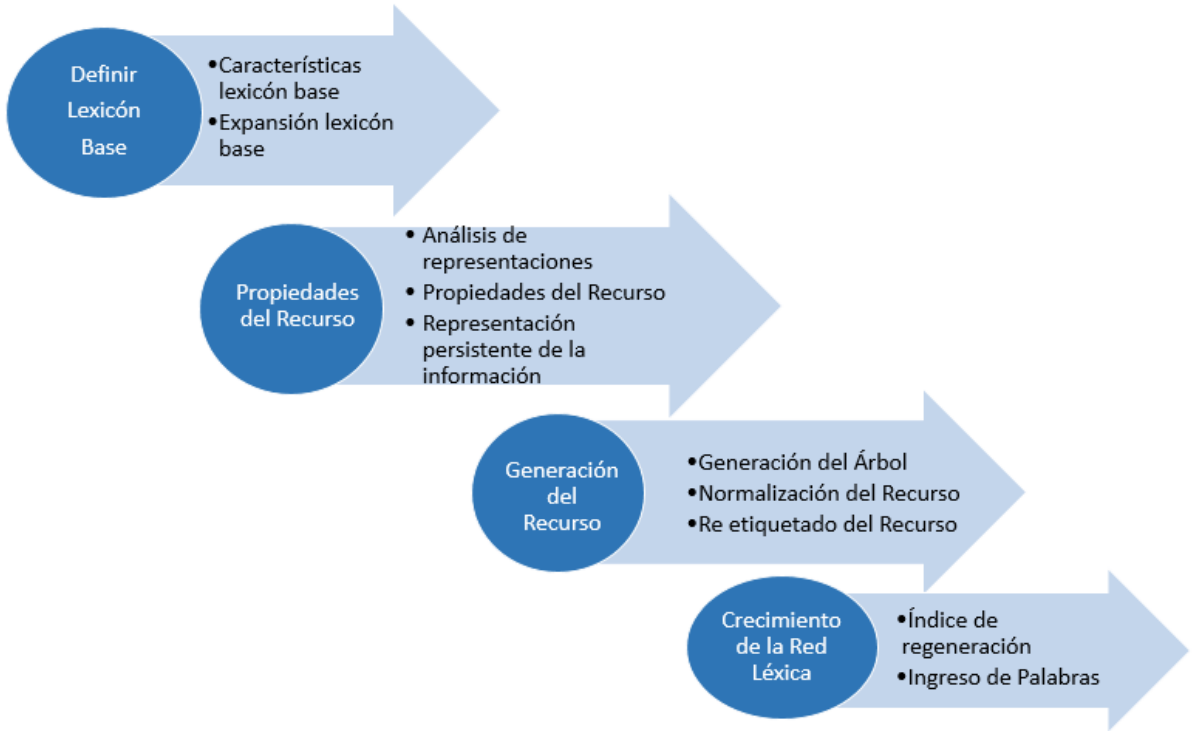



Figura 4-1 Esquema de construcción del lexicón afectivo expandido

- Definir lexicón base: En esta etapa se mencionan las características que debe tener el lexicón base y cuál fue el proceso de expansión
- Propiedades del Recurso: En esta etapa se realiza un análisis de las representaciones, la representación seleccionada y su definición, además de enumerar las propiedades del recurso.
- Generación del recurso: En esta etapa se genera el recurso léxico en una estructura de árbol, se realizan normalizaciones del recurso y se define el proceso de corrección de intensidad a través de encuesta.
- Crecimiento de la Red léxica: Se define el índice para la regeneración de la red léxica y el proceso para incorporación de nuevas palabras.

Una vez creado el recurso y definida su representación se realiza la evaluación que permite comprobar la hipótesis.

## 4.1 Definir Lexicón Base

### 4.1.1 Características lexicón base

El lexicón a seleccionar debe incorporar una traducción al español o poseer una versión multilenguaje, y por otra parte debe estar etiquetado para las ocho emociones de Plutchick [14]. Entre los lexicones que cumplen con estas características se encuentran [9] y [13]. El primero contiene palabras que son afectivas más las palabras que ocurren en comentarios etiquetados en alguna emoción. El segundo lexicón incluye la intensidad basado en 6 métricas de similitud, de las cuales se seleccionaron aquellas que reportan mejores y más consistentes resultados, específicamente la métrica *Path* [54] basada en estructura y la métrica de Resnik [55] basada en contenido de información (IC). En [13] el lexicón propuesto toma como base EmoLEX[15] pero selecciona las palabras que poseen al menos un synset (agrupación de sinónimos de WordNet [15]) con antecesor *emotion* o *feeling* es decir palabras con raíz afectiva. Lo anterior se debe a que el resultado de la similaridad será relevante sí y solo sí, los synset de las palabras poseen emocionalidad. En la Figura 4-2 se representa un extracto de la representación del árbol de hiperónimos en WordNet de 4 palabras pertenecientes a las clases *anger*: *rage*, *ire* y *malice*. Como sustantivos, cada una de estas palabras poseen 5, 2 y 2 *synset* respectivamente. Tal como se observa, es muy distinto analizar la similaridad entre la palabra *rage.synset1* (*rage.n.5*) e *ire.synset2 (ire.n.2*) o entre las palabras rage.n.3 con *ire.n.1*.

Si una palabra de la clase es similar a la palabra más intensa de la clase entonces se asume que ésta también es intensa, ahora bien, si la palabra tiene baja o nula similitud con la palabra más intensa se asume que posee baja intensidad emocional en la clase. Tal como se presenta en la Figura 4-3, en la clase alegría, la palabra con mayor intensidad es éxtasis. La similitud entre éxtasis y la palabra “deleite”, “cosquilla” y “diversión” es de 0.16, 0.06 y 0.14 respectivamente. Basado en estos resultados se obtiene que cuando se expresa la palabra “cosquillas” está palabra es menos intensa o expresa alegría en menor intensidad que la palabra “deleite”.

[15] Lexicón en inglés con su asociación a ocho emociones básicas de Plutchik, y dos sentimientos positivo y negativo. Obtenido a través de anotaciones manuales mediante *Crowdsoursing*.

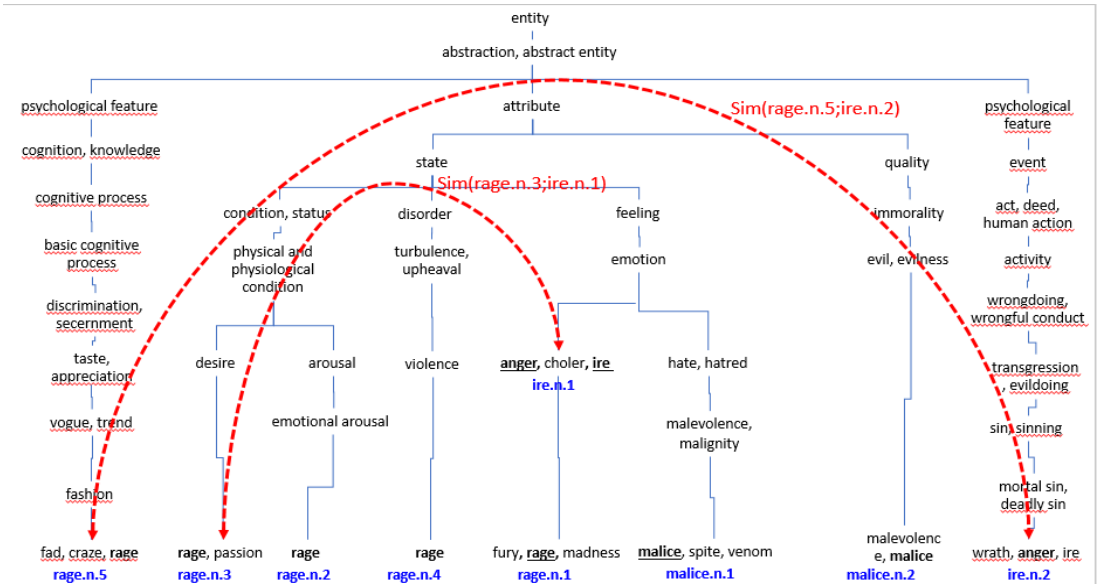



Figura 4-2 Ejemplo de extracción de sinónimos de la palabra *rage*-ira desde *WordNet* recorriendo los *synset*.


Figura 4-3 Representación de la similitud entre éxtasis y las palabras deleite, cosquilla y diversión pertenecientes a la clase alegría.

### 4.1.2 Expansión

El lexicón seleccionado es IntensiLEX [13]. Este recurso fue seleccionado ya que está etiquetada la intensidad de las palabras en la emoción, a diferencia de EmoLEX que solo las etiqueta sin agregar la intensidad. Por otra parte este proyecto da continuidad al trabajo IntensiLEX [13] desarrollado dentro de una línea de trabajo del grupo de investigación. IntensiLEX posee 640 palabras emotivas en inglés (con su traducción al español) etiquetadas con intensidad (231 palabras únicas), cuya distribución por clase se grafica en la Figura 4-4, como cantidad sin expandir (palabras).

Estas 640 palabras son expandidas con los sinónimos disponibles en WordNet, de este proceso se agregaron en promedio 5 sinónimos por cada clase-palabra, específicamente 3526 sinónimos lo que deja un total de 4166 palabras. Las palabras obtenidas por clase después de expandir el lexicón se presenta en Figura 4-4, como cantidad expandido(palabras). El crecimiento de las clases fue un 670% y se incorporaron 440 palabras en promedio, donde la clase que más palabras incorporó fue *fear* y la que menos incorporó fue *trust,* con 576 y 325 palabras respectivamente.

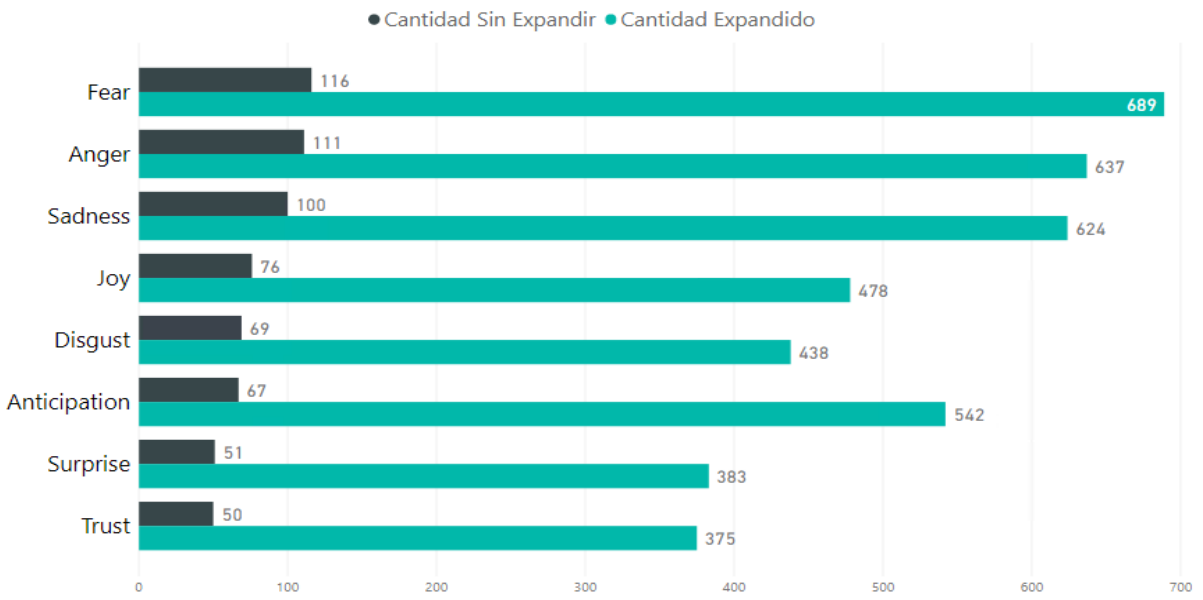



Figura 4-4 Frecuencia de palabras por clase lexicón sin expandir y expandido.

Cabe destacar que al incorporar los sinónimos se validó que todos ellos provengan de la clase *emotion* o *feeling* de la base de conocimiento *WordNet,* similar a la forma en que extrae *WordNet-Affect* [51] su taxonomía desde W*ordNet*. En la Figura 4-5 se muestra un extracto de la taxonomía de WordNet-Affect.

Hasta esta etapa el lexicón se encuentra en inglés, aunque cada palabra posee una traducción al español (obtenida desde *EmoLEX*). Algunos sinónimos ya pertenecían al lexicón inicial y los restantes se les incorporó su traducción utilizando Google Translate, tal como se hizo en el lexicón original *EmoLEX*. Considerando que una palabra puede tener más de un *synset* con padre *feeling* o *emotion*, se seleccionó la tupla palabra-*synset* cuya intensidad fuese mayor por cada clase. Un ejemplo de extracción de sinónimos desde *WordNet de la palabra rage*(ira) y su respectiva navegación a través de los respectivos *synset,* se muestra en la Figura 4-2.

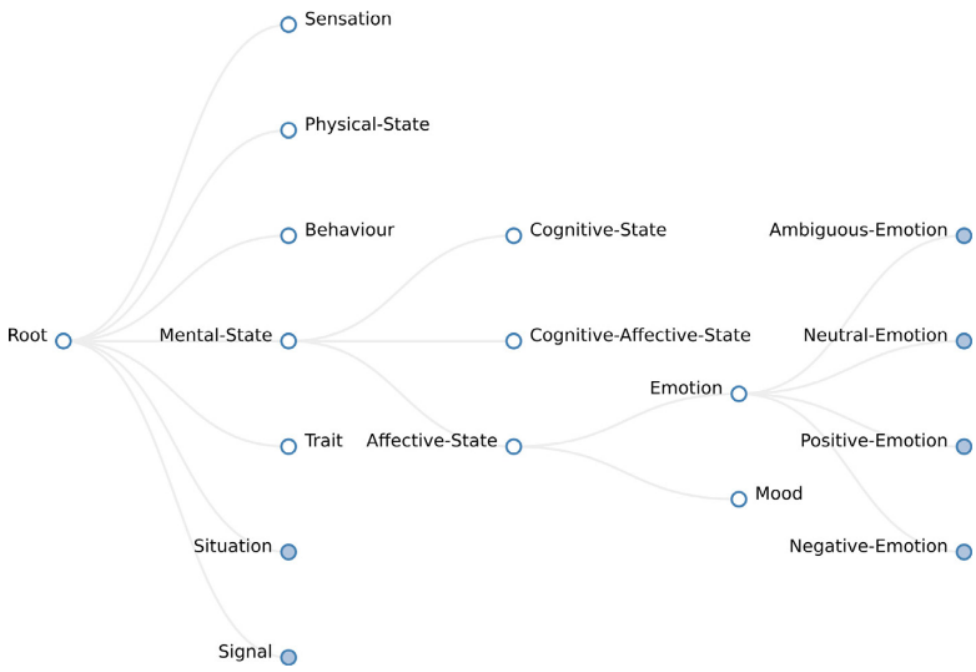



Figura 4-5 Extracto de la taxonomía de *WordNet-Affect [48]*.

Cada una de las palabras en español es expandida utilizando un diccionario online. En la extracción se obtuvieron todos los *synset* de la palabra con sus sinónimos. Lo anterior fue implementado utilizando una herramienta creada con la API TECO, disponible en https://dsi.face.ubiobio.cl/somos/#tools.

Dado que el objetivo de esta Tesis es proveer un lexicón afectivo en español etiquetado con intensidad y clasificado en las ocho emociones de Plutchik, fue necesario analizar las traducciones provistas con el propósito de seleccionar las que correspondían a palabras o expresiones literales que no son utilizadas en español. Estas palabras fueron reemplazadas por una palabra o expresión más común y representativa en español. Para esta tarea se requirió la participación de un evaluador quien analizó las 425 palabras en inglés con su traducción al español, cabe destacar la diferencia entre el tamaño del lexicón 640 y las 425 palabras analizadas se explica ya que existen tuplas(palabra+traducción) que se encuentran etiquetadas en más de 1 clase. En total fueron corregidas o reemplazadas 113 palabras.

En resumen, el lexicón cuenta 1159 palabras únicas en español y 4166 clase-palabra totales para el lexicón expandido. Para etiquetar la intensidad de las nuevas palabras, se calculó la similitud utilizando la métrica *Path* [54].

Para desplegar la intensidad a las palabras en español agregadas al lexicón se consideró:

- Asignar la intensidad de la palabra del lexicón en inglés desde donde fue extendida.
- Si existen palabras que son sinónimos de más de una palabra del lexicón en inglés, y que además posee distintos valores de intensidad se discrimina generando un promedio simple de la intensidad, un ejemplo de esto se muestra en la tabla Tabla 4-1, donde se ve la extracción de la emoción *fear*, obteniendose las palabras iniciales *quiver*, *shudder*, *shiver* y *thrill*, su traducción para todas las palabras iniciales del inglés es estremecerse y la palabra del lexicón es atemorizar y en todas representan una intensidad distinta, en este caso se aplica el promedio simple y 0.46 será el valor de intensidad de la palabra atemorizar de la clase *fear*.

| **Inglés** | | | **Español** | | |
|---|---|---|---|---|---|
| **Emoción** | **Palabra inicial** | | **Interpretación** | **Palabra del lexicón** | **Intensidad** |
| ***Fear*** | *quiver* | | estremecerse | atemorizar | 0.64 |
| | *shudder* | | estremecerse | atemorizar | 0.64 |
| | *shiver* | | estremecerse | atemorizar | 0.28 |
| | *thrill* | | estremecerse | atemorizar | 0.28 |

Tabla 4-1 Sinónimos desde el inglés

Luego con el lexicón traducido, revisado y evaluado tenemos 956 palabras que se encuentran más de una vez en la tupla *clase+palabra* con distintas intensidades. En la Tabla 4-2 se muestran como ejemplos las intensidades dela palabra afán y espíritu que se encuentran en las ocho clases afectivas.

Que una palabra se etiquete en las ocho clases afectivas producirá demasiada ambigüedad en el análisis de emociones. Por lo tanto, para depurar el lexicón se eliminaron las palabras que están en más de 4 clases y que tengan las menores intensidades, en el caso de que la intensidad menor sea la misma para más de 4 clase se eliminan todas éstas. En el ejemplo la palabra afán se clasifica en las 8 emociones y de acuerdo a lo anterior se seleccionan las mejores 4 intensidades, es decir, sorpresa con 0.25, confianza con 0.2, anticipación con 0.2 y la última con 0.1666667. No obstante, existen 5 emociones con este último valor de intensidad, por lo cual sólo se consideran en el lexicón las 3 primeras clases. Así también para el caso de la palabra espíritu en donde en la clase tristeza tiene 0.16666667, luego para confianza y anticipación su valor es 0.2, pero aún está en cinco emociones por lo tanto se selecciona el siguiente valor que es 0.25 de las clases sorpresa, miedo, ira y aversión, por lo tanto, espíritu solo pertenecerá a la clase alegría.

Terminado el proceso de eliminación el lexicón queda con 3212 palabras y su distribución final se puede ver en la Figura 4-6.

| Clase | Clase inglés | Palabra | Intensidad *Path* |
|---|---|---|---|
| ira | *anger* | afán | 0.16666667 |
| aversión | *disgust* | afán | 0.16666667 |
| miedo | *fear* | afán | 0.16666667 |
| alegría | *joy* | afán | 0.16666667 |
| tristeza | *sadness* | afán | 0.16666667 |
| anticipación | *anticipation* | afán | 0.2 |
| confianza | *trust* | afán | 0.2 |
| sorpresa | *surprise* | afán | 0.25 |
| tristeza | *sadness* | espíritu | 0.16666667 |
| confianza | *trust* | espíritu | 0.2 |
| anticipación | *anticipation* | espíritu | 0.2 |
| sorpresa | *surprise* | espíritu | 0.25 |
| miedo | *fear* | espíritu | 0.25 |
| ira | *anger* | espíritu | 0.25 |
| aversión | *disgust* | espíritu | 0.25 |
| alegría | *joy* | espíritu | 0.5 |

Tabla 4-2 Palabras del lexicón con distintas intensidades para la misma palabra en distintas clases afectiva.

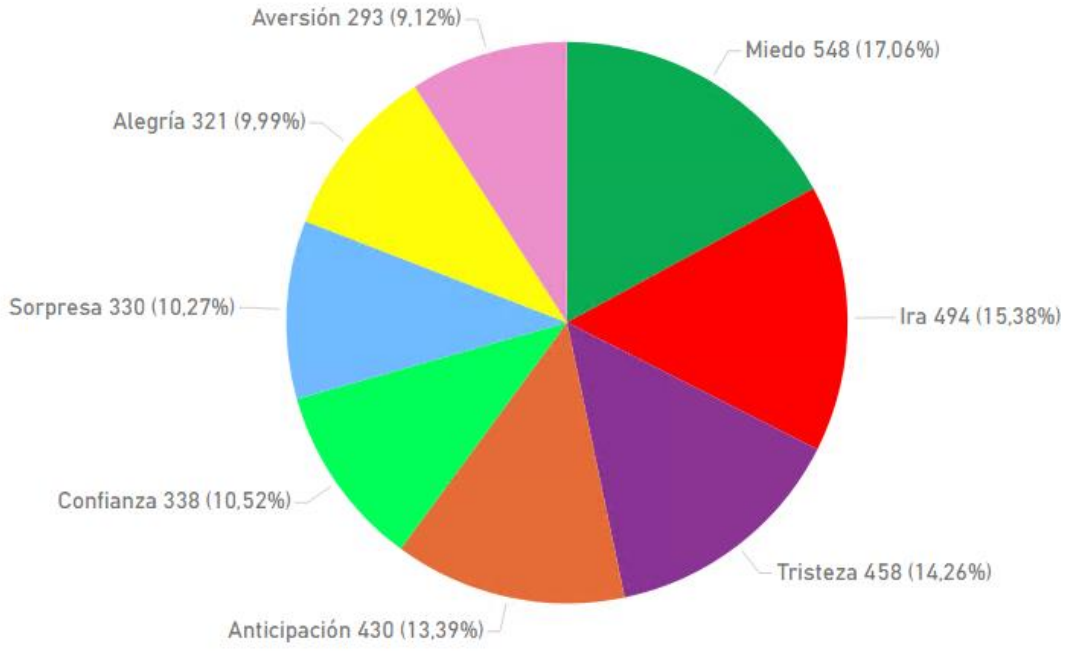



Figura 4-6 Distribución final por clase afectiva del lexicón expandido

La distribución por clase afectiva queda de la siguiente forma, ver Figura 4-7.

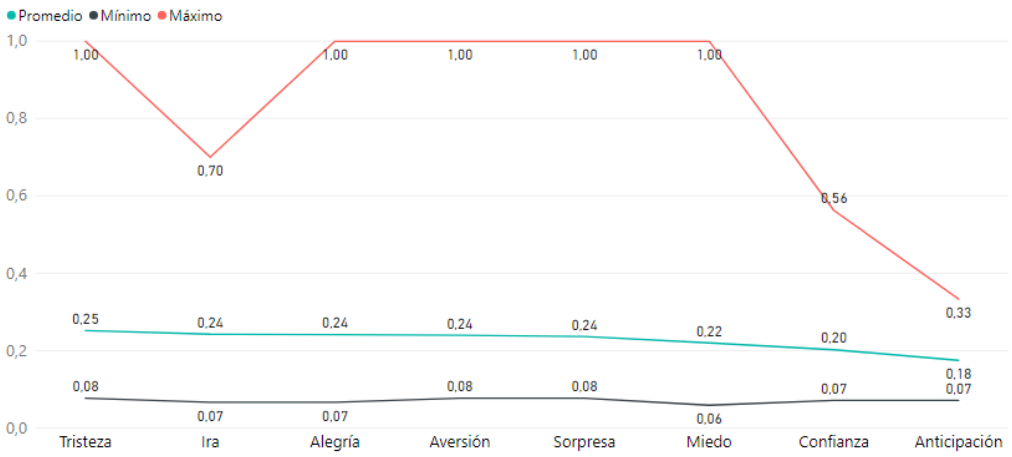



Figura 4-7 Gráfico de clase afectiva y sus valores de intensidad mínima, máxima y promedio.

En este punto ya se ha definido el recurso léxico afectivo en español etiquetado en las ocho emociones de *Plutchik* y con su intensidad basado en la métrica de similaridad de *Path.* El recurso cuenta con 1159 palabra únicas en español, para un total de 4166 palabras totales. Cabe destacar que RedPal solo representa palabras emotivas, es decir solo contiene palabras con carga afectiva o emocional. Ejemplo: Este ha sido el peor día de trabajo, esta frase solo la palabra “peor” tiene carga afectiva.

Esta es la primera versión RedPal, sobre la cual se definirán funciones de normalización y algoritmos para generación, expansión y crecimiento.

## 4.2 Propiedades del Recurso

### 4.2.1 Análisis de representaciones

A continuación, se describen las representaciones analizadas para la propuesta.

Se analizaron las relaciones *IS_A* y de sinonimia entre las palabras del lexicón según *WordNet*. Para ello, se construyeron grafos dirigidos y no dirigidos sobre cada clase afectiva de *Plutchik*.

- Una primera representación revisada corresponde a un grafo dirigido sobre la clase afectiva *anger* y sus sinónimos obtenidos desde *WordNet,* en donde los nodos representan las palabras y las relaciones su relación de sinonimia, la dirección del enlace representa la relación X es sinónimo de Y, ver Figura 4-8. En esta

representación, la clase afectiva de *Plutchik*, *anger* y *rage* el cual es el sinónimo más intenso *en WordNet* para anger*,* en donde los *synset* iniciales serán los nodos hijos de los nodos centrales *anger y rage*, y luego se incorporan los sinónimos de cada una de palabras navegando a través de los *synsets* que tengan raíces *emotion* o *feeling*.

- Una segunda representación revisada del lexicón es su representación jerárquica de árbol basada en el nivel de intensidad afectiva encontrada y distribuida en niveles, ver Figura 4-9. En esta representación la clase *anger* se construyó basado en su intensidad, donde furia está en el segundo nivel jerárquico, luego furia y rabia en el tercer nivel, luego molestia, rozar y vejación dependiente de furia en un cuarto nivel jerárquico, y así hasta completar todas las jerarquías. Notar que furia esta repetido esto es debido que esta representación se hace antes de realizar todas las operaciones en el lexicón que se mencionan en los capítulos siguientes.

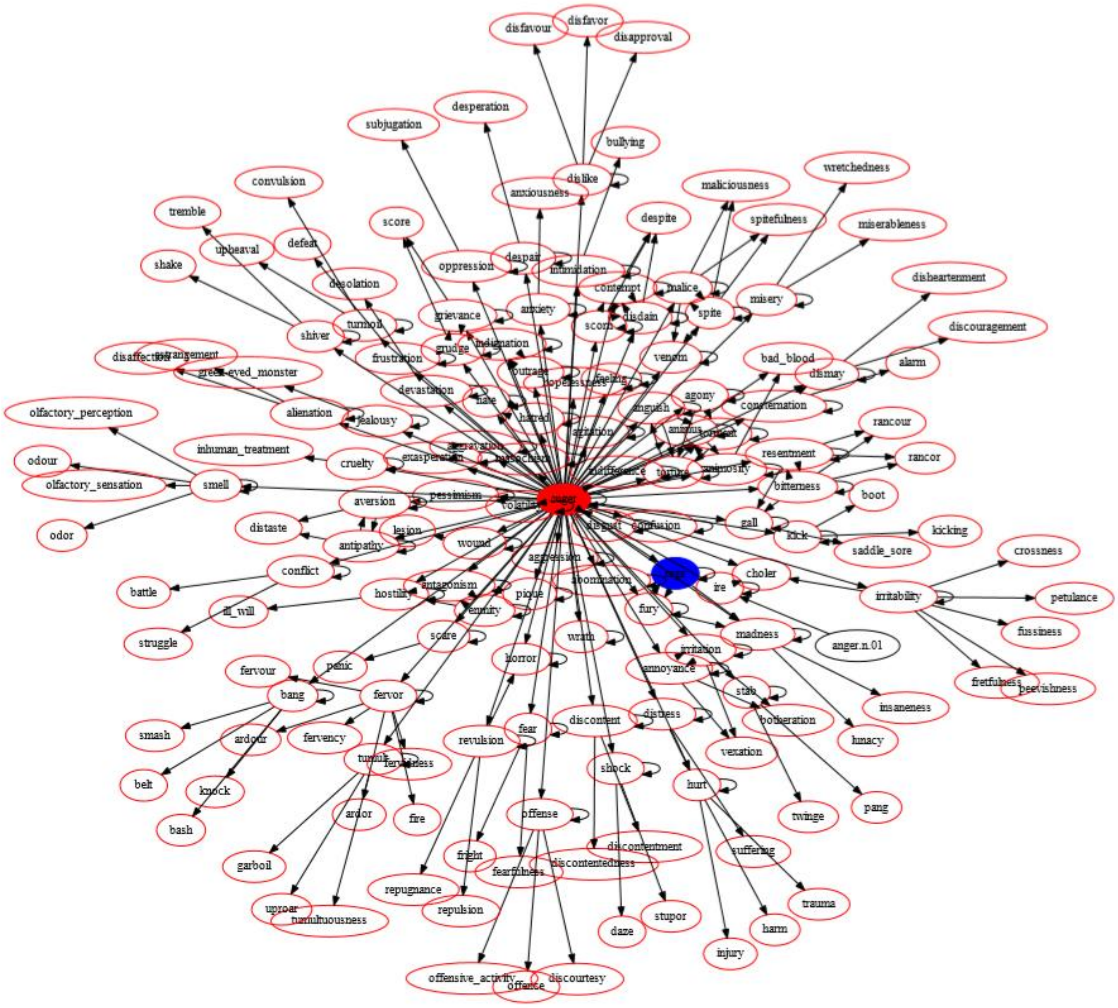



Figura 4-8 Ejemplo de una de las representación analizada, clase afectiva *anger [14]* y su palabra de referencia *rage*.

Como aspecto positivo de la primera representación revisada, posee la estructura léxica, pero por el contrario no es posible adicionar características ni permite asignar jerarquías.

La segunda representación revisada permite incorporar la jerarquía y atributos, además esta representación permite hacer un manejo más eficiente de las intensidades de las palabras de una clase, expansión, actualización y uso.

Cabe recordar que en las ciencias de la computación la diferencia entre un árbol y un grafo, es que en el árbol dos vértices están conectados exactamente por un camino.

Concordante con la literatura [29] [32] [34], el recurso léxico más extendido *WordNet* [15] y las teorías lexicográficas [11] [35], la mejor forma de representar un recurso léxico que éste enfocado en el desempeño, la organización, y que permita operaciones semánticas y sintácticas, es una estructura de árbol jerárquico.

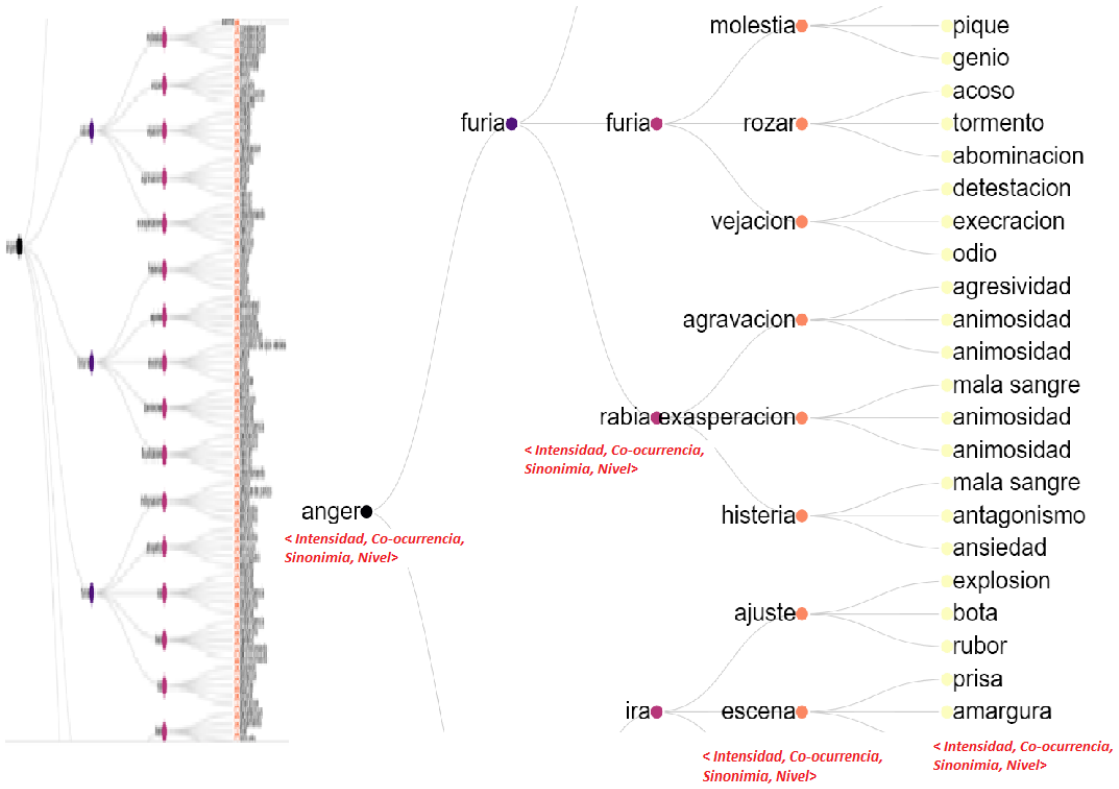



Figura 4-9 Porción de la representación distribuida en 5 niveles, para la clase afectiva *Anger*.

### 4.2.2 Propiedades del recurso

En este apartado se definen las propiedades y características relevantes del recurso léxico.

- El recurso léxico propuesto contará con atributos mínimos: la clase afectiva basada en Plutchik, la palabra extraída del lexicón expandido y su intensidad basada en *Path.*
- El recurso léxico representará la relación inicial *IS_A* derivada de la extracción inicial realizada desde *WordNet,* que representa la sinonimia.
- Co-ocurrencia: Cuando palabra Y existe en la frase X, genera una emoción o intensidad diferente a la palabra por si sola. Dado que el recurso solo contiene palabras emotivas estos atributos serán utilizados para procesar y luego identificar expresiones afectivas.
- Frases coloquiales: Incorporación de palabras que no están reconocidas en los diccionarios tradicionales y que corresponden a expresiones propias de una cultura, pero se pueden relacionar a una palabra y clase afectiva, y por lo tanto heredar su intensidad, Ej. chilenismos.

### 4.2.3 Representación persistente de la información

Con el objeto de controlar las propiedades y atributos del recurso léxico se define una base de datos relacional para la representación del recurso léxico, su modelo de datos físico se muestra en la Figura 4-10.

La base de datos está compuesta por 5 tablas:

- **Clase_afectiva**: Almacena las clases afectivas que dispondrá el recurso léxico, por ahora tendrá las 8 clases afectivas de Plutchik.
- **Palabra**: Se utiliza para almacenar las palabras del recurso léxico.
- **Padre**: Se utiliza para almacenar la estructura de árbol del recurso léxico indicando el padre, para hacer referencia a la palabra.
- **Tipo_relacion**: Almacena las relaciones que dispondrá el recurso léxico.
- **Relacion**: Almacena la relación que representa la palabra en el recurso léxico, en la etapa inicial todas representan la relación es_un (*is_a)*.

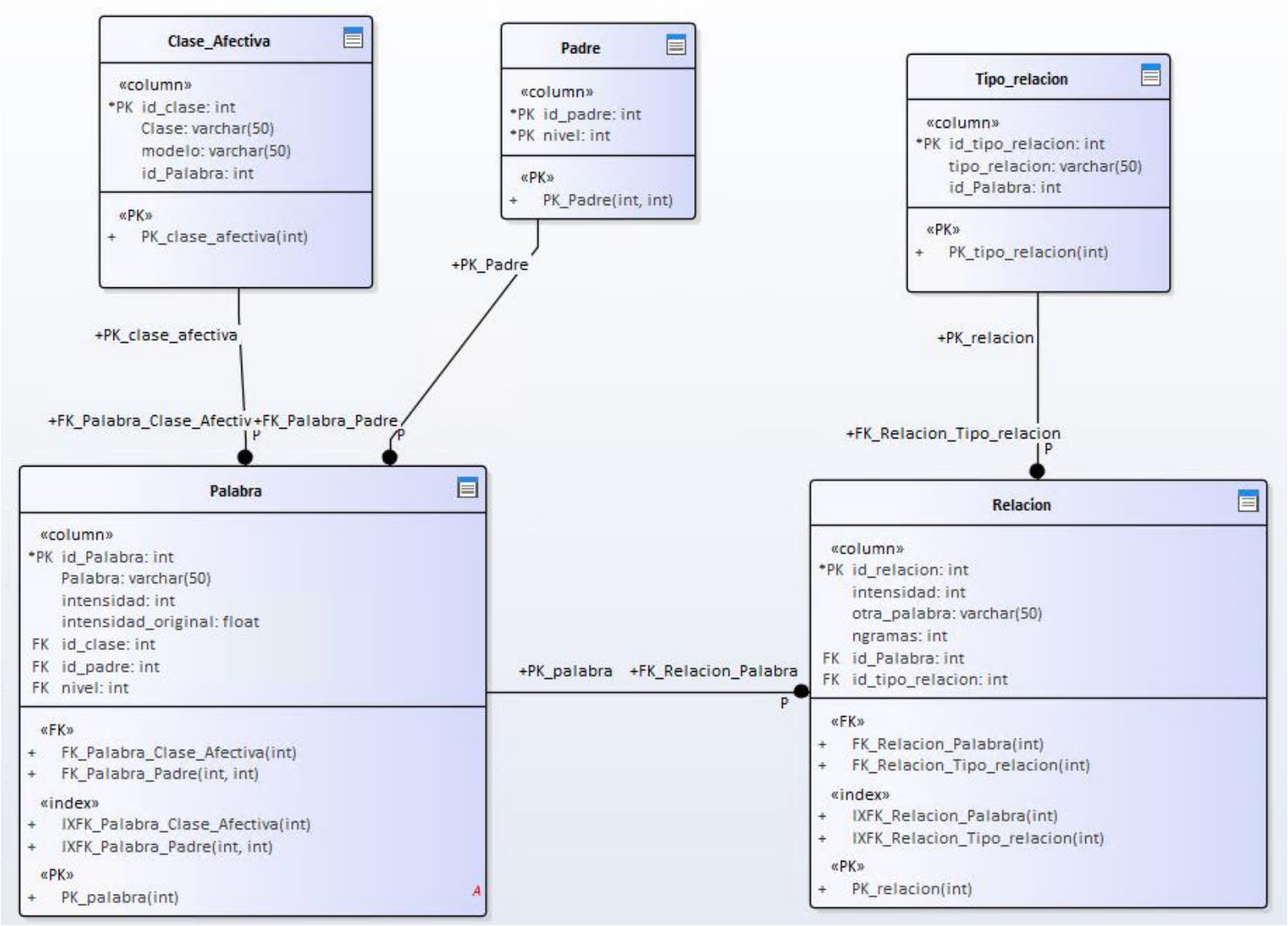



Figura 4-10 Diseño base de datos RedPal.

## 4.3 Generación del recurso

### 4.3.1 Generación árbol

Antes de realizar la generación del árbol verificamos la distribución de los datos por clase afectiva. En donde encontramos valores muy dispersos no permitiendo encontrar agrupaciones marcadas y mucha concentración de la tupla “clase afectiva-palabra” a un mismo valor de intensidad. Lo anterior se muestra en el análisis para las clases: Alegría (Figura 4-11), Anticipación (Figura 4-12), Aversión (Figura 4-13).

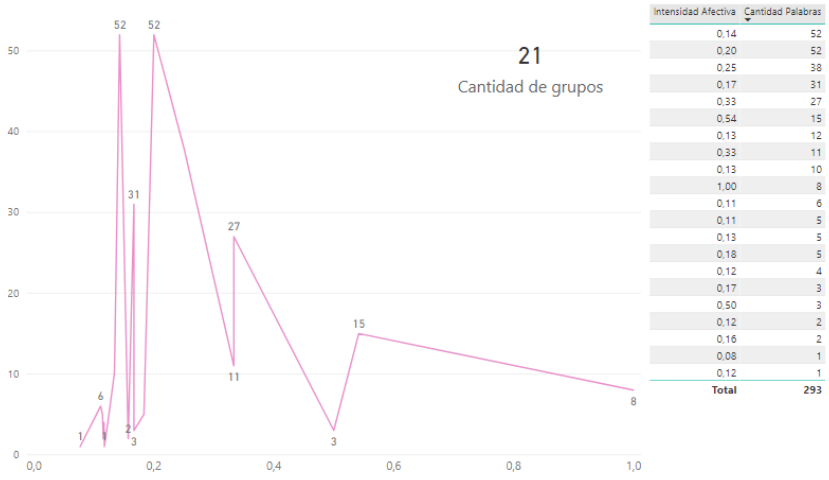

| Intensidad Afectiva | Cantidad Palabras |
|---|---|
| 0,17 | 69 |
| 0,20 | 52 |
| 0,14 | 34 |
| 1,00 | 26 |
| 0,25 | 24 |
| 0,13 | 16 |
| 0,20 | 16 |
| 0,20 | 15 |
| 0,22 | 15 |
| 0,08 | 6 |
| 0,14 | 6 |
| 0,14 | 6 |
| 0,16 | 6 |
| 0,14 | 5 |
| 0,14 | 5 |
| 0,07 | 3 |
| 0,12 | 3 |
| 0,14 | 3 |
| 0,17 | 3 |
| 0,18 | 3 |
| 0,09 | 1 |
| 0,11 | 1 |
| 0,13 | 1 |
| 0,14 | 1 |
| 0,31 | 1 |
| **Total** | **321** |

Figura 4-11 Análisis de intensidad apara Alegría (*Joy*).


| Intensidad Afectiva | Cantidad Palabras |
|---|---|
| 0,20 | 121 |
| 0,17 | 119 |
| 0,14 | 52 |
| 0,25 | 25 |
| 0,21 | 20 |
| 0,09 | 19 |
| 0,15 | 12 |
| 0,16 | 11 |
| 0,08 | 7 |
| 0,09 | 7 |
| 0,16 | 7 |
| 0,33 | 6 |
| 0,14 | 5 |
| 0,17 | 5 |
| 0,11 | 4 |
| 0,07 | 3 |
| 0,33 | 3 |
| 0,08 | 1 |
| 0,13 | 1 |
| 0,17 | 1 |
| 0,18 | 1 |
| **Total** | **430** |

Figura 4-12 Análisis de intensidad apara Anticipación (*Anticipation*).

De las figuras Figura 4-11, Figura 4-12 y Figura 4-13 podemos extraer la cantidad de agrupaciones que tienen las clases afectivas según si intensidad inicial derivada de *WordNet* y la cantidad de palabras asociadas a esa intensidad.




| Intensidad Afectiva | Cantidad Palabras |
|---|---|
| 0,14 | 52 |
| 0,20 | 52 |
| 0,25 | 38 |
| 0,17 | 31 |
| 0,33 | 27 |
| 0,54 | 15 |
| 0,13 | 12 |
| 0,33 | 11 |
| 0,13 | 10 |
| 1,00 | 8 |
| 0,11 | 6 |
| 0,11 | 5 |
| 0,13 | 5 |
| 0,18 | 5 |
| 0,12 | 4 |
| 0,17 | 3 |
| 0,50 | 3 |
| 0,12 | 2 |
| 0,16 | 2 |
| 0,08 | 1 |
| 0,12 | 1 |
| **Total** | **293** |

Figura 4-13 Análisis de intensidad para Aversión (*Disgust*).

Por lo tanto, se realiza una construcción inicial del árbol jerárquico basado en intensidad, no balanceado y respetando que a un mismo valor de intensidad las palabras se encuentren en un mismo nivel jerárquico. El algoritmo para la generación inicial está en la Figura 4-14, este algoritmo tiene complejidad computacional *O ( n Log n ).*

***Generación Árbol Inicial por Intensidad:***

*Por cada clase afectiva:(Alegría, Anticipación, Aversión, Confianza, Ira, Miedo, Sorpresa, tristeza)*

> ***Lista:*** *Ordenar por* ***Valor*** *de intensidad en orden descendiente y calcular el número de elementos para ese* ***Valor***
>
> *Asignar como nodo inicial a la clase afectiva sin padre*
>
> > *Recorrer la Lista y verifica la cantidad de elementos que correspondan al orden*
> >
> > > ***Distribuir:*** *Realizar calculo según la cantidad de elementos para el* ***Valor*** *y distribuir según la cantidad de padres*
> > >
> > > *Asignar los elementos al nivel y padre correspondiente*
> > >
> > > *Verificar que los elementos asignados no superen a la cantidad de elementos para el* ***Valor***
> > >
> > > *Verificar la cantidad de hijos por nivel correspondan al valor de* ***Distribuir***

Figura 4-14 Algoritmo para la generación del árbol inicial basado en la intensidad afectiva.

Para la clase aversión(*disgust*) el algoritmo realizaría lo siguiente, ordena los valores por su intensidad en orden descendente donde esta tiene 8 elementos con el valor máximo 1 los cuales serán hijos de la clase aversión, luego pasa al siguiente nivel por intensidad 0.541667 donde existen 15 palabras con ese valor de intensidad y se deben distribuir según los 8 padres disponibles asignando 2 hijos a cada palabra que tenía la intensidad 1 y así hasta asignar la última palabra con intensidad 0.076923.

En la tabla Tabla 4-3 se resume la información sobre la altura y nodos para cada clase afectiva del recurso léxico, y en la Tabla 4-4 ejemplos de palabras que tienen más de 5 nodos hijos:

| Clase Afectiva | Altura | Nodos |
|---|---|---|
| Miedo | 28 | 548 |
| Tristeza | 22 | 458 |
| Ira | 28 | 494 |
| Anticipación | 21 | 430 |
| Alegría | 25 | 321 |
| Aversión | 21 | 293 |
| Confianza | 23 | 338 |
| Sorpresa | 24 | 330 |

Tabla 4-3 Información de árbol inicial.

| Clase Afectiva | Palabra Padre | Cantidad de nodos hijos |
|---|---|---|
| Alegría | Fuego | 6 |
| Alegría | Albricias | 5 |
| Ira | Desesperación | 7 |
| Ira | Impío | 6 |
| Tristeza | Compungido | 8 |
| Tristeza | Desgracia | 7 |
| Confianza | Elogio | 8 |
| Confianza | Dulce | 8 |
| Miedo | Sensibilidad | 8 |
| Miedo | Temor | 8 |
| Sorpresa | Incitación | 7 |
| Sorpresa | Sumisión | 6 |
| Anticipación | Desencantamiento | 7 |
| Anticipación | Desear | 7 |
| Aversión | Venganza | 8 |
| Aversión | Lesión | 8 |

Tabla 4-4 Ejemplos de los nodos que tengan más de cinco nodos hoja.

Para verificar la eficiencia del árbol inicial revisaremos fórmulas generales de árboles *K-ario* [56]*:*

$$h = [log_k(k-1) + log_k(numero\_nodos) - 1]$$

Ecuación 4-1 Altura estimada de un árbol *k-ario*.

$$numero\_nodos = \frac{k^{h+1} - 1}{k - 1}$$

Ecuación 4-2 Número estimado de nodos de un árbol k-ario.

Aplicando la Ecuación 4-2, k corresponde al número máximo de hijos por nodo y h corresponde a la altura del árbol *k-ario.* Aplicando la ecuación anterior podemos obtener una estimación teórica de la altura del árbol, ver Tabla 4-5.

| **Clase Afectiva** | **Cantidad Palabras** | **h mínima(k=2)** | **h mínima(k=3)** | **h mínima(k=4)** |
|---|---|---|---|---|
| Miedo | 548 | 9 | 6 | 5 |
| Tristeza | 458 | 8 | 6 | 5 |
| Ira | 494 | 8 | 6 | 5 |
| Anticipación | 430 | 8 | 6 | 5 |
| Alegría | 321 | 8 | 5 | 4 |
| Aversión | 293 | 8 | 5 | 4 |
| Confianza | 338 | 8 | 5 | 4 |
| Sorpresa | 330 | 8 | 5 | 4 |

Tabla 4-5 Estimación teórica de altura de árbol k-ario.

Aplicando la Ecuación 4-2, k corresponde al número máximo de hojas por nodo y *h* corresponde a la altura del árbol *k-ario.* Aplicando la ecuación anterior podemos obtener una estimación teórica del número máximo de nodos para el árbol *k-ario* y la altura mínima, ver Tabla 4-6.

| Clase Afectiva | h mínima (k=2) | Máximo Elementos (k=2) | h mínima (k=3) | Máximo Elementos (k=3) | h mínima (k=4) | Máximo Elementos (k=4) |
|---|---|---|---|---|---|---|
| Miedo | 9 | 1023 | 6 | 1093 | 5 | 1395 |
| Tristeza | 8 | 511 | 6 | 1093 | 5 | 1395 |
| Ira | 8 | 511 | 6 | 1093 | 5 | 1395 |
| Anticipación | 8 | 511 | 6 | 1093 | 5 | 1395 |
| Alegría | 8 | 511 | 5 | 364 | 4 | 341 |
| Aversión | 8 | 511 | 5 | 364 | 4 | 341 |
| Confianza | 8 | 511 | 5 | 364 | 4 | 341 |
| Sorpresa | 8 | 511 | 5 | 364 | 4 | 341 |

Tabla 4-6 Estimación teórica del número máximo de nodos para un árbol k-ario y la altura mínima.

Comparando los valores teóricos con los obtenidos, Tabla 4-6, en la generación inicial del árbol podemos indicar que está muy lejos de valores teóricos aceptables por lo se deben definir parámetros para normalización de intensidades y reorganización del árbol.

### 4.3.2 Normalización del recurso

De la tabla Tabla 4-7 podemos también notar que las clases afectivas ira, confianza y anticipación su valor máximo de intensidad no es 1, los valores mostrados en la tabla y gráfico están redondeados a 2 decimales. Este aspecto se debe considerar antes realizar la normalización de los valores máximos y sus respectivas distribuciones para las clases afectivas que su valor máximo no es uno (1), luego de las desambiguaciones realizadas y el valor extraído desde *WordNet*. En la Tabla 4-7, vemos que para la clase Ira su valor máximo es (0.7), para la clase confianza su valor máximo es (0.56) y para la clase anticipación su valor máximo es (0.33). Esto se debe a que en la extracción desde el inglés una palabra tenía más de un sinónimo, y en estos casos se seleccionó el promedio simple de la intensidad de estas palabras, es decir la palabra en promedio expresa solo una intensidad en la clase.

| Clase Afectiva | Palabras | Promedio | Mínimo | Máximo |
|---|---|---|---|---|
| Tristeza | 458 | 0,25 | 0,08 | 1,00 |
| Ira | 494 | 0,24 | 0,07 | 0,70 |
| Alegría | 321 | 0,24 | 0,07 | 1,00 |
| Aversión | 293 | 0,24 | 0,08 | 1,00 |
| Sorpresa | 330 | 0,24 | 0,08 | 1,00 |
| Miedo | 548 | 0,22 | 0,06 | 1,00 |
| Confianza | 338 | 0,20 | 0,07 | 0,56 |
| Anticipación | 430 | 0,18 | 0,07 | 0,33 |
| **Total** | **3212** | **0,23** | **0,06** | **1,00** |

Tabla 4-7 Distribución por clase afectiva y sus valores de intensidad promedio, mínima y máxima.

Se debe verificar que el valor máximo de nivel superior sea homogéneo en todas las clases afectivas, es decir todas las clases afectivas deben tener límite superior el valor 1, la fórmula para normalizar el valor del *Path* en la Ecuación 4-3.

$$PathNormalizado = (NuevoPath_{n-1} - AntiguoPath_{n-1}) + \ Path_n$$

Ecuación 4-3 Path Normalizado

Donde el *Path_normalizado* normalizará a 1 como valor máximo y mantendrá la diferencia entre los valores siguientes, para la clase ira su valor máximo de es 0.70 el cual al normalizar será 1. El valor de la categoría dos es 0.5 y tiene de diferencia 0.2 con la categoría 1(0.7-0.5), si la categoría es uno el valor de la categoría dos es 0.8, el proceso para las primeras 4 categorías de la clase ira está en la Tabla 4-8. Este proceso que se repetirá para las clases ira, confianza y anticipación, cuyo valor máximo no es uno

| Clase | Antigua Intensidad | Nueva Intensidad Normalizada | Nueva Intensidad | Nivel en el árbol |
|---|---|---|---|---|
| Ira | 0,70 | 1,00 | 3333 | 1 |
| Ira | 0,50 | 0,80 | 3309 | 2 |
| Ira | 0,40 | 0,70 | 3296 | 3 |
| Ira | 0,33 | 0,63 | 3288 | 4 |

Tabla 4-8 Normalización de las primeras 4 categorías de Ira.

Finalizada la extracción y validación del lexicón, la normalización de intensidades afectivas, podemos prescindir de los valores extraídos de *WordNet* y podemos llevarlos a una escala propia que permita trabajar de mejor forma el árbol y clasificador, eliminando valores difíciles de leer, conversión escala intensidad se explica según Ecuación 4-4, Ecuación 4-5 y Ecuación 4-6.

$$Suma\_intensidades = \sum_{i=1}^{n} I_i$$

Ecuación 4-4 Ecuación para la suma de intensidades del lexicón.

$$Total\_Elementos = \sum_{i=1}^{n} 1$$

Ecuación 4-5 Ecuación para el conteo de elementos del lexicón.

Para las ecuaciones Ecuación 4-4 y Ecuación 4-5:

- $\mathbf{i}$ representa al índice del elemento.
- $\mathbf{I}$: representa la intensidad afectiva.
- $\mathbf{n}$: representa la cantidad de elementos del lexicón.

$$Nueva\_intensidad = \frac{(Suma\_intensidades + intesidad\_orig\_elemento) * Total\_Elementos}{(Suma\_intensidades - intensidad\_orig\_elemento)}$$

Ecuación 4-6 Ecuación para el cálculo de la nueva intensidad.

Al aplicar la formula la podemos ver en Figura 4-15 los datos para todas las clases afectivas, su antigua intensidad y la nueva intensidad, el nivel del árbol en que están las palabras y la cantidad de palabras para el nivel.

| Clase | Antigua Intensidad | Nueva Intensidad Normalizada | Nueva Intensidad | Nivel en el árbol |
|---|---|---|---|---|
| Anticipación | 0,33 | 1,00 | 3333 | 1 |
| Confianza | 0,56 | 1,00 | 3333 | 1 |
| Ira | 0,70 | 1,00 | 3333 | 1 |
| Alegría | 1,00 | 1,00 | 3333 | 1 |
| Aversión | 1,00 | 1,00 | 3333 | 1 |
| Miedo | 1,00 | 1,00 | 3333 | 1 |
| Sorpresa | 1,00 | 1,00 | 3333 | 1 |
| Tristeza | 1,00 | 1,00 | 3333 | 1 |
| Anticipación | 0,25 | 0,92 | 3323 | 2 |
| Alegría | 0,31 | 0,31 | 3249 | 2 |
| Miedo | 0,50 | 0,50 | 3272 | 2 |
| Sorpresa | 0,50 | 0,50 | 3272 | 2 |
| Tristeza | 0,50 | 0,50 | 3272 | 2 |
| Ira | 0,50 | 0,80 | 3309 | 2 |
| Confianza | 0,50 | 0,94 | 3326 | 2 |
| Aversión | 0,54 | 0,54 | 3277 | 2 |
| Anticipación | 0,21 | 0,88 | 3318 | 3 |
| Alegría | 0,25 | 0,25 | 3242 | 3 |
| Tristeza | 0,31 | 0,31 | 3250 | 3 |
| Sorpresa | 0,33 | 0,33 | 3252 | 3 |
| Miedo | 0,36 | 0,36 | 3255 | 3 |
| Ira | 0,40 | 0,70 | 3296 | 3 |
| Confianza | 0,47 | 0,90 | 3321 | 3 |
| Aversión | 0,50 | 0,50 | 3272 | 3 |
| Anticipación | 0,20 | 0,87 | 3317 | 4 |
| Alegría | 0,22 | 0,22 | 3238 | 4 |
| Tristeza | 0,25 | 0,25 | 3242 | 4 |
| Sorpresa | 0,27 | 0,27 | 3244 | 4 |
| Aversión | 0,33 | 0,33 | 3252 | 4 |
| Miedo | 0,33 | 0,33 | 3252 | 4 |
| Ira | 0,33 | 0,63 | 3288 | 4 |
| Confianza | 0,33 | 0,77 | 3305 | 4 |

Figura 4-15 Ejemplos de nuevas intensidades para las clases afectivas, sus primeros 4 niveles.

Luego de aplicar la fórmula de nueva intensidad tenemos un valor máximo normalizado para todas las clases de 3333 y luego se distribuyen según su intensidad normalizada.

La implementación final del recurso estará disponible en el sitio de recursos del grupo de investigación SoMoS, en la Figura 4-16 se puede ver la visualización del Árbol de RedPal.

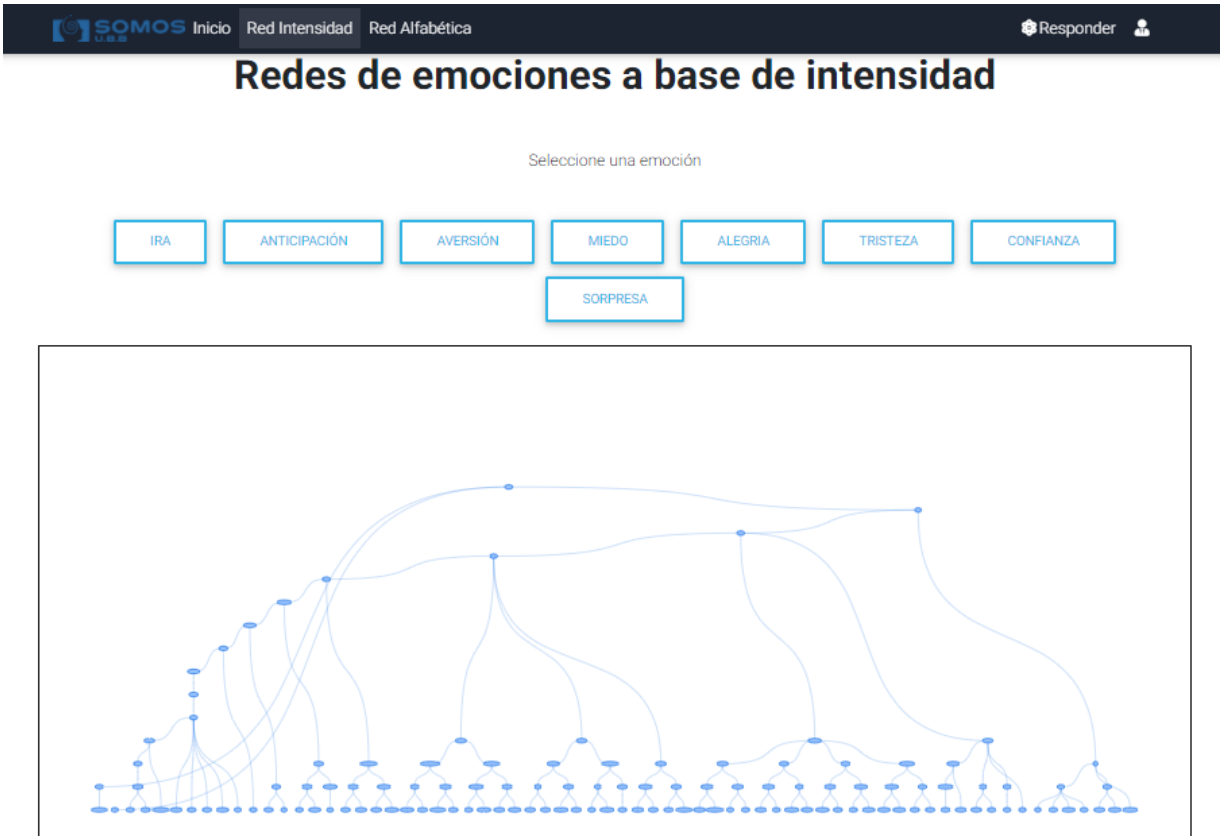



Figura 4-16 Árbol inicial RedPal, captura de pantalla de red intensidad *[57]*

### 4.3.3 Re etiquetado del recurso

El re etiquetado se realizará a través de encuestas las cuales serán aplicadas a los usuarios de la herramienta web de la red léxica afectiva en español, basado en una estrategia crowdsourcing. A continuación, se definen los criterios con los cuales seleccionar las palabras que serán consultadas, lo anterior dado que el árbol esta ordenado y los resultados de las encuestas deben permitir ser interpretados para regenerar el nuevo orden de intensidad.

Casos para generar encuestas:

- Cuando el padre tiene solo un hijo: Se genera solo una encuesta con el padre y su hijo.

- Cuando el padre tiene dos hijos: Se genera solo una encuesta con el padre y su hijo de mayor intensidad y su hijo de menor intensidad.
- Rama: Se generaran tomando un nodo pivote y se realizará una navegación dos niveles hacia arriba y dos niveles hacia abajo, hacia arriba el padre y hacia el padre de su padre, y hacia abajo uno de sus hijos y uno de los hijos de su hijo, un ejemplo se puede ver en la Figura 4-17, donde el nodo pivote es alienación su navegación hacia arriba abarca a los nodos aborrecimiento y a pesar de, y su navegación hacia abajo abarca los nodos desdén y sentimiento de culpa.

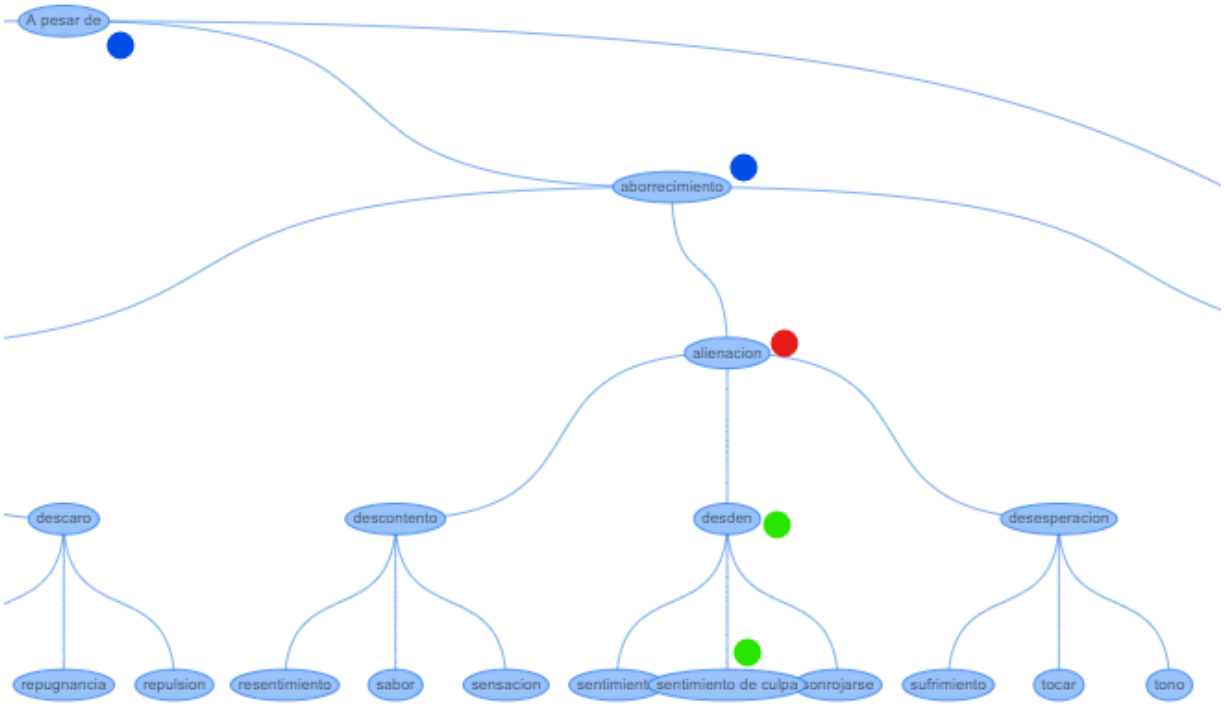



Figura 4-17 Generación de una encuestas por rama.

La encuesta consistirá en solicitar al navegante de la plataforma ordenar de mayor a menor, las palabras clasificadas dentro de una emoción, en la Figura 4-18 podemos ver la presentación que tendrá la encuesta para ser respondida.

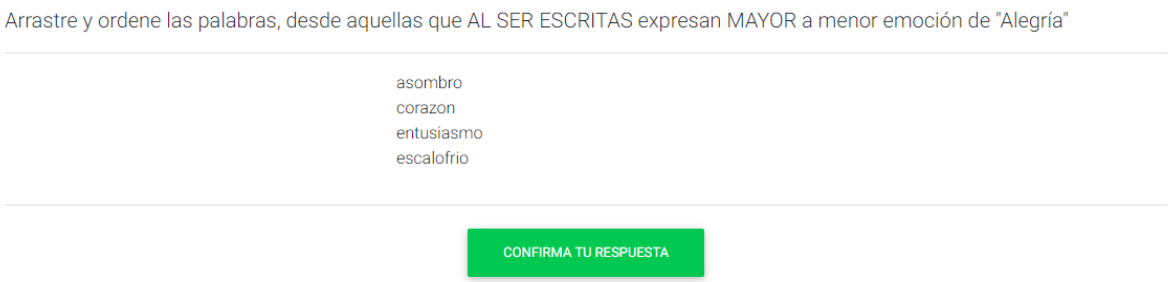



Figura 4-18 Presentación de la encuesta.

Para mejorar la confiabilidad de la intensidad consultada y analizar el nivel de acuerdo en las encuestas, se sugiere al menos una encuesta sea contestada tres veces.

## 4.4 Crecimiento de la Red Léxica

La red léxica podrá crecer de dos formas a través de ingreso directo de palabras a la plataforma o a través de palabras candidatas extraídas por el clasificador.

### 4.4.1 Índice de Regeneración

La actualización del recurso léxico se basará en un índice informativo, este índice representará un valor porcentual qué representa el porcentaje de completitud del árbol k-ario. A partir de este valor se determina regenerar el árbol e incorporar la información recopilada a través del sistema de encuestas. La función de regeneración estará disponible en cualquier momento para ser ejecutada, en base a la cantidad de nodos de las clase afectivas, se propone un valor de 70% para la regeneración, ver índice para regeneración en

Figura 4-19 y Ecuación 4-7, la complejidad computacional de este algoritmo es de O ( *n log n* ).

```
Índice Regeneración:

Entradas: (RedPal)

     Seleccionar Árbol (Clase Afectiva)

     Parámetros: Altura, extrae la altura del árbol desde
los parámetros de RedPal

          Recorrer Árbol (Clase Afectiva): verifica si
          existen hojas libres

               Acumular_hojas_libres (Clase Afectiva, Nivel,
          Hojas libres) Acumula hojas libres por nivel,
          registrar en parámetros RedPal

               Acumular_Hojas_total(Clase Afectiva) Acumular
          total de hojas del árbol


          Calcular Índice (Clase Afectiva): Verificar% de
          completitud del árbol

               AHTA = Acumular_Hojas_total(Clase Afectiva)

               Verificar_max_altura(): comparar con max
               elementos-kario(altura)

               IREG= calcular el índice de regeneración total
```

Figura 4-19 Algoritmo índice de regeneración

$$IREG = \frac{AHTA}{numero\_nodos}\ x\ 100$$

Ecuación 4-7 Ecuación cálculo de índice de regeneración.

Donde *AHTA,* representa la cantidad de hojas por clase afectiva

### 4.4.2 Incorporación de palabras a RedPal

La incorporación de palabras deberá a lo menos contar con criterios mínimos de aceptación que serán:

- Clase Afectiva, es decir la palabra a incorporar debe estar clasificada en al menos una emoción.
- Palabra hermana o más cercana, en intensidad.
- Palabra nueva

Ejemplo: si quiere agregar la palabra: enfado, se necesita además la clase afectiva a la cual pertenece en este caso Ira y la palabra hermana impaciencia y se localizaría la palabra más cercana en el árbol Figura 4-20.

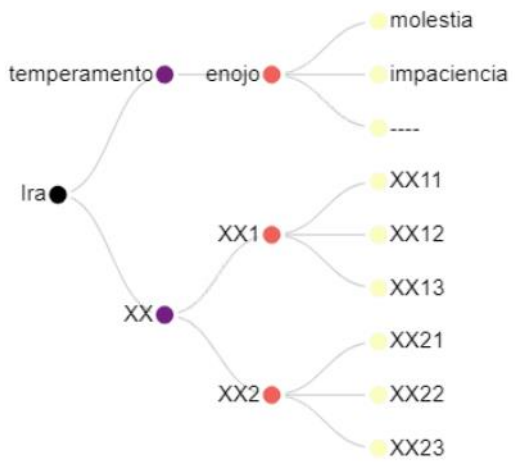



Figura 4-20 Extracto de árbol de Ira, antes de la incorporación de nueva palabra.

Aplicando el algoritmo descrito en la Figura 4-21 para la incorporación directa de nuevas palabras, se tiene el siguiente árbol resultante Figura 4-22, la intensidad asignada a enfado es el promedio de los hermanos hojas en este caso el promedio de la intensidad de molestia e impaciencia. El algoritmo para incorporación de una nueva palabra tiene una complejidad computacional de *O (n log n).*

```
Incorporación Nueva Palabra:

Entradas: (Palabra Nueva, Clase Afectiva,
Palabra hermana)

      Seleccionar Árbol (Clase Afectiva)

      Buscar en el árbol (Palabra Hermana)

            Verificar Rama (): verifica si
            existen hojas libres

            Si existen hojas libres: Insertar
            Palabra Nueva en la hoja Libre.

            CalcularIntensidad(): promedio de los
            hermanos hoja

            De lo contrario: Regenerar Árbol
            (Palabra Nueva): regenerar el Árbol
            incorporando la palabra nueva
```

Figura 4-21 Algoritmo para la incorporación de una nueva palabra.

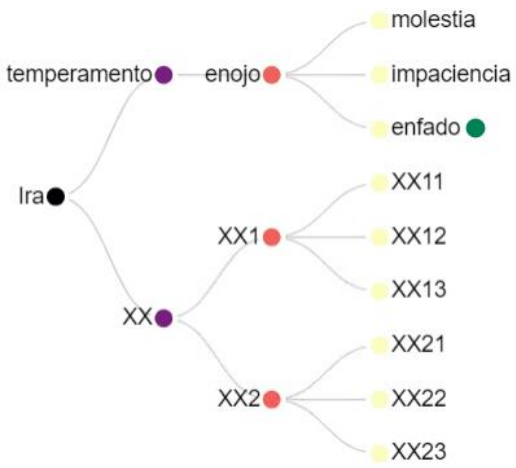



Figura 4-22 Extracto de árbol de Ira, *después* de la incorporación de nueva palabra ”Rubor”.

### 4.4.3 Agregar desde Clasificador

En cada análisis de emociones desde un corpus o frase el recurso léxico, se verificará que las palabras no contenidas en el recurso léxico, serán incorporadas a una estructura temporal de RedPal, tomando como clase afectiva para la palabra o palabras la clase predominante en la clasificación. Este proceso debe pasar por un análisis detallado, puesto que la red solo contiene palabras afectivas, por lo tanto, es necesario determinar el rol de la palabra en la frase y la intensión que esta tiene. Por ejemplo, una palabra que juega el rol de adjetivo es probable que posea una interpretación afectiva, no obstante, las herramientas de procesamiento de lenguaje natural en español no son 100% efectivas en su identificación, ver algoritmo en Figura 4-23, este algoritmo tiene complejidad computacional *O (n).*

```
Incorporación Nuevo Sustantivo Clasificador:
Revisión clasificadora(texto)
      Si encuentra sustantivo ()
            Retornar agregar sustantivo(sustantivo, clase
predominante)
Agregar_sustantivo (sustantivo, clase predominante)
Entradas: sustantivo clase predominante
      Seleccionar Árbol (clase predominante)
      Buscar en el árbol (Palabra Hermana)
            Si existe retornar error_clasificador ()
            No existe:
                  Dejar el sustantivo en estructura temporal para
                  realizar análisis exhaustivo de incorporación
```

Figura 4-23 Algoritmo para incorporar palabra desde clasificador.

## 4.5 Propiedades de RedPal

En la tabla Tabla 4-9 se describen las características que hacen superior a RedPal por sobre lexicones de trabajos anteriores, su creciente potencial de expansión y desambiguación.

### 4.5.1 RedPal v/s Lexicones

| Características | RedPal | Lexicón |
|---|---|---|
| Idioma | Español | Muy pocos |
| Basado en intensidad afectiva | Si | Muy pocos |
| Crecimiento: Incorporación de palabras | Constante | Estático Ej: EmoLEX[16] |
| Actualización: modificación de la representación | Constante | Estático |
| Regeneración | Constante | Estático |
| Retroalimentación: Crowdsoursing | A través de encuestas | NO |
| Exportar | Posibilidad de exportar, el contenido de RadPal: en distintos formatos, texto plano, Json, cvs. Para su análisis y reutilización | Sólo un formato, principalmente texto plano. |
| Clasificador | El clasificador entregara un resumen, de cada análisis para revisión y/o análisis posteriores(json) | No existe posibilidad. |

Tabla 4-9 Características de RedPal v/s lexicones

## 4.6 Conclusiones del Capítulo

En este capítulo se analizó todo lo referente a la representación del recurso léxico RedPal, las representaciones disponibles en la literatura, las principales características y restricciones. Con esta información se decide representar en una estructura jerárquica de árbol y soportado por un modelo de datos relacional y los algoritmos para construir la primera versión del recurso léxico. El recurso léxico contempla desambiguación constante a través de encuestas, lo que desencadenará en regeneración del recurso. El crecimiento constante es otra característica soportado por operaciones de incorporación directa y a través del clasificador de nuevas palabras. Además, una comparación de características entre RedPal y las características de los lexicones presentados en el capítulo tres, podemos destacar las características positivas

[16] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

de RedPal por sobre los lexicones vistos en la revisión bibliográfica, en donde podemos destacar la actualización y expansión constante, desambiguación y retroalimentación.

# 5 Experimento de evaluación

En esta sección se describe el diseño de un experimento de evaluación de RedPal, propuesto en este trabajo.

El experimento de evaluación consistirá en tres evaluaciones:

1. Analizar RedPal en su versión inicial.
2. Analizar RedPal en versiones con regeneración e incorporación de nuevas palabras.
3. Análisis hibrido de agresividad

## 5.1 Analizar rendimiento del análisis de emociones utilizando RedPal

### 5.1.1 Herramienta

Para llevar a cabo esta evaluación, se implementó una herramienta en Python que usa librerías propias de éste lenguaje y funciones de implementación propia, que conforman el recurso léxico RedPal. Esta herramienta implementa las siguientes funciones:

- Tokenización: Esta etapa consiste en descomponer un elemento del corpus (frase) en distintas partes.
- *Lower*: Eliminar las mayúsculas en las palabras ya tokenizadas.
- Puntuación: Eliminar caracteres de puntuación (coma, punto, punto y coma, etc.)
- *Stopwords*: Eliminar *stopwords* en español existentes en el texto.
- *Skip-grams* [58]: Además de análisis de palabras el clasificador, permitirá analizar frases compuestas, co-ocurrencia o expresiones coloquiales que existan en el clasificador, por lo cual se realizará *skip-gram* de la frase o corpus, basado el máximo n-gram existente en RedPal, para realizar posteriormente el análisis de afectos. En la Figura 5-1 se puede ver el ejemplo *skip-gram,* en donde para un n=2, realiza una agrupación de dos palabras consecutivas en el texto, y para un n=3 realiza el mismo proceso anterior pero con tres palabras consecutivas.

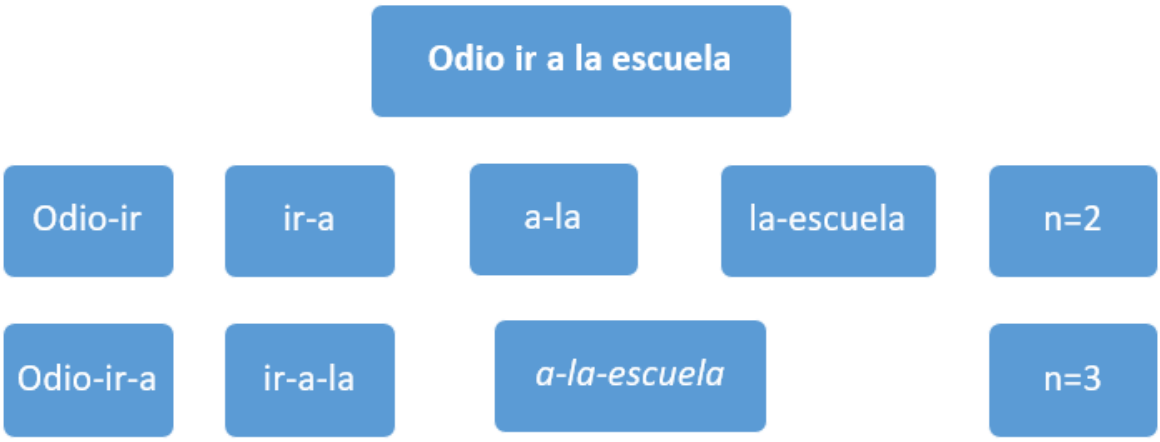



Figura 5-1 Ejemplo de *skip-gram*

- Análisis: El análisis evalúa la coincidencia del texto o corpus, aplicando las funciones antes mencionadas, acumulando las intensidades por clase afectiva.
- Resultados: El clasificador entrega como respuesta del análisis, un detalle de las clases afectivas con la cantidad de palabras encontradas para la clase, la intensidad acumulada. También tiene una opción de depuración donde se entregará el análisis de todas las clases afectivas y el detalle de todas las palabras afectadas.

### 5.1.2 Métricas de Evaluación

La métrica utilizada para analizar el desempeño de los clasificadores en el análisis de emociones es Recall, la que es igual a la porción de documentos de una clase que son clasificados correctamente, en función de los casos de esa misma clase, correcta e incorrectamente clasificados. Se obtendrán 2 *Recall* para evaluar RedPal.

*Recall* Hit: Corresponderá a la porción de evaluaciones cuando el clasificador encontró una y sólo una clase relevante, esto corresponde cuando la suma de intensidades solo tiene una clase mayor.

Para ellos se considera los siguientes casos:

- True Positives (*tp*): elementos a los que el clasificador asignó una clase relevante y ésta correcta.
- False Positives (*fp*): elementos a los que el clasificador asignó una clase relevante y esta no era correcta.

- False Negatives (*fn*): elementos a los que el clasificador asignó la clase no relevante y esta no era correcta.
- True Negatives (*tn*): elementos a los que el clasificador asignó la clase no relevante y esta era correcta.

*Recall* Ambiguo: Corresponderá a la porción de evaluaciones cuando el clasificador encontró más de una clase relevante, esto corresponde cuando la suma de intensidades no puede asignar solo una clase mayor.

Para ellos se considera los siguientes casos:

- True Positives (*tp*): elementos a los que el clasificador asigno la clase relevante y esta correcta.
- False Positives (*fp*): elementos a los que el clasificador asigno la clase relevante y esta no era correcta.
- False Negatives (*fn*): elementos a los que el clasificador asigno la clase no relevante y esta no era correcta.
- True Negatives (*tn*): elementos a los que el clasificador asigno la clase no relevante y esta era correcta.

### 5.1.3 Factores de Análisis

- **Cobertura**: Corresponde a la cantidad de palabras que pudieron ser analizadas en el texto sin considerar los *stopwords*. En la Figura 5-2 se muestra un ejemplo en el que la cobertura es del 50%, dos palabras para analizar y solo una afectiva. Es importante recordar que RedPal solo contienen palabras afectivas, y los comentarios o frases a analizar contienen cualquiera de las palabras de un idioma, no todas las palabras a analizar son afectivas.
- **Frases clasificadas:** Corresponde una frase que pertenece al corpus y que al realizar el análisis obtiene una cobertura mayor a 0, y por lo tanto se pudo clasificar en una o más clases afectivas.
- **Clase relevante:** Corresponde a identificar con el análisis la misma clase afectiva etiquetada en el corpus.

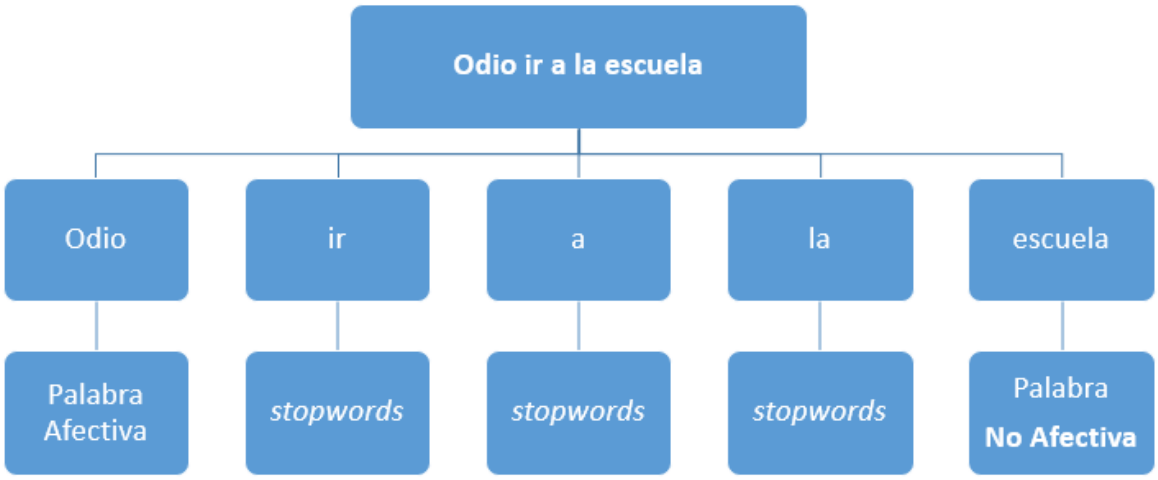



Figura 5-2 Análisis de frase para calcular cobertura.

### 5.1.4 Corpus

El corpus utilizado en la evaluación contiene 3012 titulares de 5 periódicos de distribución en Chile [59] y que esta etiquetado en las ocho clases afectivas. De los 3017 titulares, 1017 están etiquetados sin afecto(sa) y 427 con ninguna clase(ninguna), por lo tanto, el universo etiquetado con las 8 clases afectivas corresponde 1568 titulares. La Figura 5-3 muestra la distribución de etiquetas en el corpus, según las ocho clases afectivas.

RedPal contiene 3212 palabras desambiguadas y revisadas, lo que corresponde a un 327% de lexicón en español (EmoLEX) el cual esta etiquetado en las ocho emociones de Plutchik, las mismas clases afectivas de RedPal. EmoLEX a pesar de estar etiquetado en las clases afectivas de RedPal no contempla la intensidad afectiva de la palabra, por lo cual se asigna un valor teórico de 100 a cada tupla clase-palabra, para que el clasificador realice la misma evaluación en ambos casos

La comparación se realizará con una versión de EmoLEX en español, que sólo contiene palabras con raíces afectivas, esta versión de EmoLEX contiene solo un 13% de las palabras de EmoLEX original.

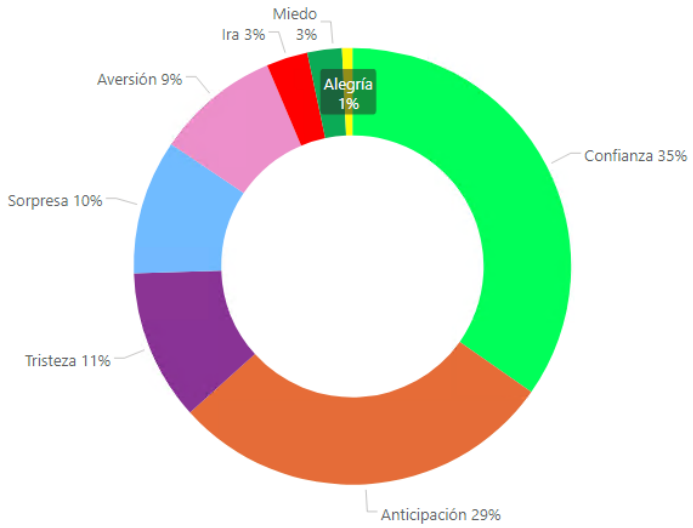



Figura 5-3 Distribución de las emociones en el corpus utilizado.

### 5.1.5 Analizar *Recall* de RedPal Inicial

El primer análisis será *Recall Hit* entre una versión inicial de RedPal, sin realizar desambiguaciones ni normalizaciones que llamaremos RedPal-V0 comparada con EmoLEX en español y una cobertura de 1% a 100% para ambos lexicones. Los resultados se resumen en la Tabla 5-1.

| Lexicón | Hits | Recall | Cobertura | Prom. Clase Relevante | Mín. Clase Relevante | Máx. Clase Relevante | Frases Clasificadas | %Hit Frase |
|---|---|---|---|---|---|---|---|---|
| EmoLex_Emo | 9 | 0,26 | 17,21 % | 1,00 | 1 | 1 | 34 | 26,47 % |
| RedPal-V0 | 45 | 0,16 | 15,71 % | 1,00 | 1 | 1 | 275 | 16,36 % |

Tabla 5-1 Recall hit para Redpal V0 y EmoLEX español.

Para el caso analizado de la Tabla 5-1, EmoLEX analizo 34 frases y Redpal-V0 275 frases, donde EmoLEX tiene un *Recall* 26% y Redpal-V0 un *Recall* de 16%. Aunque ambos *Recall* son bajos, podemos decir que mientras EmoLEX hace hit (*tp*) a un 26,47% de las frases analizadas RedPal-V0 hit (*tp*) a un 16,36% de las frases analizadas, las coberturas promedio son 17,21% para EmoLEX y 15,71% para RedPal-V0. Podemos destacar que RedPal-V0 hace un 500% más de hit *(tp)* que EmoLex y analiza con solo una clase afectiva relevante un 808% más frases.

En las siguientes figuras podemos ver la distribución por clase afectiva de los hits *(tp),* en la Figura 5-4 se muestra la distribución para RedPal-V0 y en la Figura 5-5 la distribución

para EmoLEX. En estas figuras podemos ver que la clase relevante para ambos casos fue confianza con un 46,67% para RedPal-V0 y un 77,78% para EmoLEX.

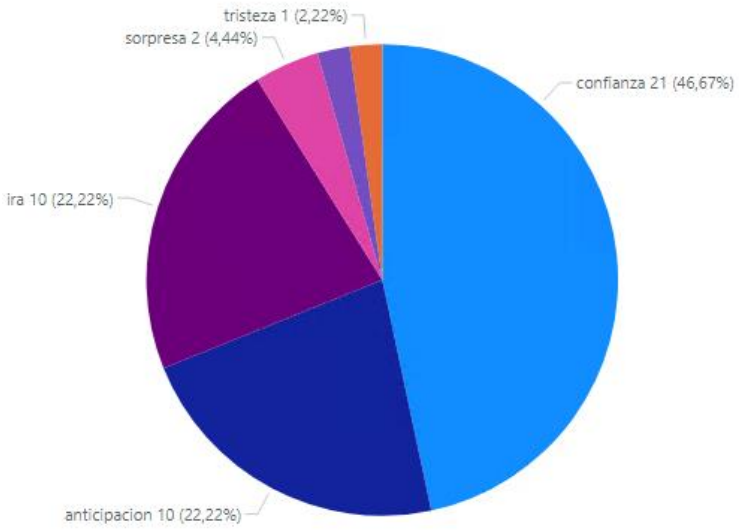



Figura 5-4 Distribución de los hits por clase afectiva para Redpal-V0

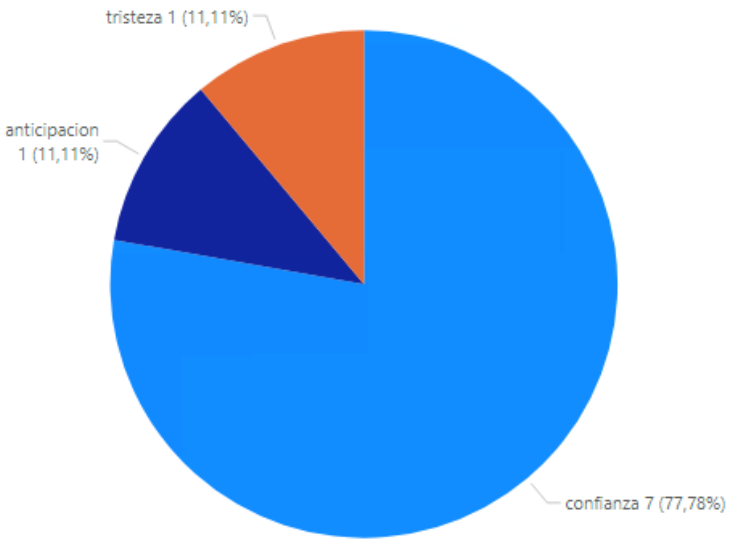



Figura 5-5 Distribución de los hits por clase afectiva para EmoLex.

El segundo análisis será *Recall* Ambiguo entre una versión inicial de RedPal, sin realizar desambiguaciones ni normalizaciones que llamaremos RedPal-V0 comparada con EmoLEX en español y una cobertura de 1% a 100% para ambos lexicones, los resultados en la

Tabla *5-2*.

| Lexicón | Hit Ambiguo | Recall | Cobertura | Prom. Clase Relevante | Mín. Clase Relevante | Máx. Clase Relevante | Frases Clasificadas | %Hit Frase |
|---|---|---|---|---|---|---|---|---|
| EmoLex_Emo | 53 | 0,58 | 14,96 % | 3,62 | 2 | 8 | 92 | 57,61 % |
| RedPal-V0 | 38 | 0,68 | 13,23 % | 2,16 | 2 | 3 | 56 | 67,86 % |

*Tabla 5-2 Recall Ambiguo para Redpal-V0 y EmoLex español*

Para el caso analizado de la

Tabla *5-2*, EmoLEX pudo analizar 92 frases y Redpal-V0 56 frases, donde EmoLEX tiene un *Recall* 58% y Redpal-V0 un *Recall* de 68%. Ambos *Recall* se consideran aceptables, podemos decir que mientras EmoLEX hace hit (*tp*) a un 57,61% de las frases analizadas RedPal-V0 hit (*tp*) a un 67,86% de las frases analizadas, las coberturas promedio son 14,96% para EmoLEX y 13,23% para RedPal-V0. Podemos desatacar que RedPal-V0 hace un 53% menos de hit *(tp)* que EmoLEX y analiza las frases con más de una clase afectiva relevante un 40% menos de frases. EmoLEX identifica en promedio 3,62 clases relevantes con una variación de dos a ocho clases relevantes, mientras RedPal-V0 tiene un promedio de 2.16 clases relevantes una variación de dos a tres clases relevantes, RedPal-V0 muestra una menor ambigüedad a identificar las clases relevantes.

En las siguientes figuras podemos ver la distribución por clase afectiva de los hits *(tp)* ambiguos, en la Figura 5-6 se muestra la distribución para RedPal-V0 y en la Figura 5-7 la distribución para EmoLEX. En estas figuras podemos ver que las clases relevantes identificadas fueron anticipación y confianza con un 71,05% para RedPal-V0 y para EmoLEX fueron alegría, anticipación, miedo, aversión, ira, tristeza y sorpresa un 22,64% para EmoLEX.

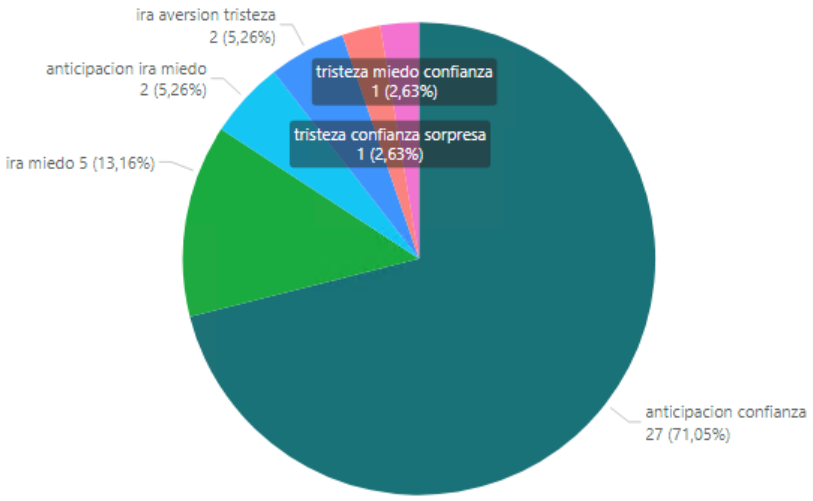



Figura 5-6 Distribución de los hits ambiguos por clase afectiva para RedPal-V0

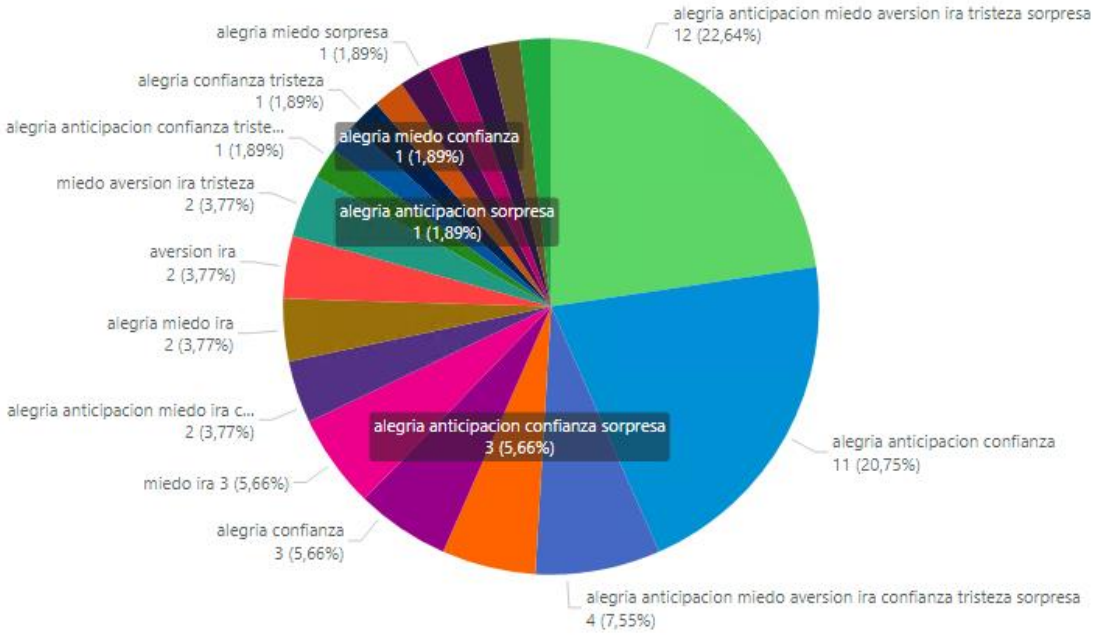



Figura 5-7 Distribución de los hits ambiguos por clase afectiva para EmoLEX.

Análisis combinado de *Recall* Hit y Ambiguo entre una versión inicial de RedPal, sin realizar desambiguaciones ni normalizaciones que llamaremos RedPa-V0 comparada con EmoLEX en español y una cobertura de 1% a 100% para ambos lexicones, los resultados en la Tabla 5-3.

| Lexicón | Hit Total | Recall Total | Cobertura | Prom. Clase Relevante | Mín. Clase Relevante | Máx. Clase Relevante | Frases Clasificadas | %Hit Frase |
|---|---|---|---|---|---|---|---|---|
| EmoLex_Emo | 62 | 0,49 | 15,57 % | 2,91 | 1 | 8 | 126 | 49,21 % |
| RedPal-V0 | 83 | 0,25 | 15,29 % | 1,20 | 1 | 3 | 331 | 25,08 % |

Tabla 5-3 Recall hit y ambiguo para Redpal-V0 y EmoLex español.

Para el análisis combinado de la Tabla 5-3, EmoLEX analizo 126 frases y Redpal-V0 331 frases, donde EmoLEX tiene un Recall 49% y Redpal-V0 un *Recall* de 25%. El *Recall* de EmoLEX se considera aceptable, para RedPal-V0 el *Recall* se considera deficiente, podemos decir que mientras EmoLEX hace hit (*tp*) a un 49,21% de las frases analizadas RedPal-V0 hit (*tp*) a un 25,08% de las frases analizadas, las coberturas promedio son 15,57% para EmoLEX y 15,29% para RedPal-V0. Podemos desatacar que RedPal-V0 hace un 20% más de hit totales *(tp)* que EmoLEX y analiza 166% más de frases. EmoLEX identifica en promedio 2,91 clases relevantes con una variación de una a ocho clases relevantes, mientras RedPal-V0 tiene un promedio de 1,20 clases relevantes una variación de una a tres clases relevantes, RedPal-V0 muestra una menor ambigüedad a identificar las clases relevantes.

### 5.1.6 Analizar *Recall* de RedPal Final

El primer análisis para la versión final de RedPal que llamaremos RedPal-VF será *Recall Hit*, esta versión contiene todo el proceso que se definieron para la creación del recurso léxico, desambiguaciones, normalizaciones y creación de la red, comparada con EmoLEX en español y una cobertura de 1% a 100% para ambos lexicones, los resultados en la Tabla 5-4.

| Lexicón | Hits | Recall | Cobertura | Prom. Clase Relevante | Mín. Clase Relevante | Máx. Clase Relevante | Frases Clasificadas | %Hit Frase |
|---|---|---|---|---|---|---|---|---|
| EmoLex_Emo | 9 | 0,26 | 17,21 % | 1,00 | 1 | 1 | 34 | 26,47 % |
| RedPal-VF | 69 | 0,21 | 15,32 % | 1,00 | 1 | 1 | 327 | 21,10 % |

Tabla 5-4 Recall hit para Redpal-VF y EmoLex español.

Para el caso analizado en la Tabla 5-4, EmoLEX analizo 34 frases y Redpal-VF 327 frases, donde EmoLEX tiene un *Recall* 26% y Redpal-VF un *Recall* de 21%. Aunque ambos *Recall* son bajos, podemos decir que mientras EmoLEX hace hit (*tp*) a un 26,47% de las frases analizadas RedPal-VF hit (*tp*) a un 21,10% de las frases analizadas, las coberturas promedio son 17,21% para EmoLEX y 15,32% para RedPal-VF. Podemos desatacar que RedPal-VF hace un casi 800% más de hit *(tp)* que EmoLEX y analiza con solo una clase afectiva relevante casi 1000% más de frases.

En las siguiente Figura 5-8 podemos ver la distribución por clase afectiva de los hits *(tp)* para RedPal-VF, y la distribución para EmoLEX ya fue presentada en la Figura 5-5. En estas figuras podemos ver que la clase relevante para Redpal-VF fue anticipación 50,72% y un 77,78% para EmoLEX.

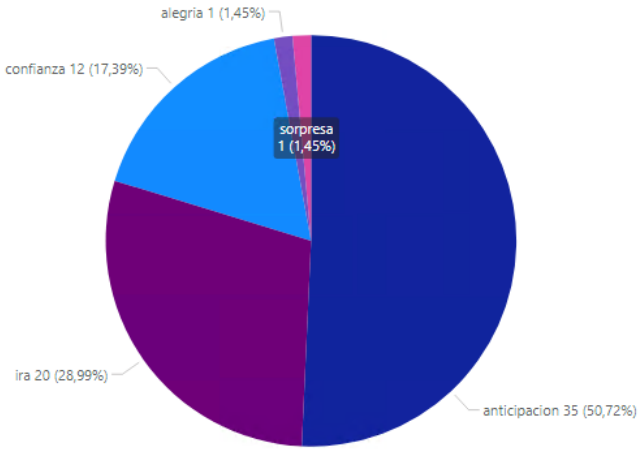



Figura 5-8 Distribución de los hits por clase afectiva para Redpal-VF

El segundo análisis para la versión final de RedPal que llamaremos RedPal-VF será *Recall Hit*, esta versión contiene todo el proceso que se definieron para la creación del recurso léxico, desambiguaciones, normalizaciones y creación de la red, comparada con EmoLEX en español y una cobertura de 1% a 100% para ambos lexicones, los resultados en la Tabla 5-5.

| Lexicón | Hit Ambiguo | Recall | Cobertura | Prom. Clase Relevante | Mín. Clase Relevante | Máx. Clase Relevante | Frases Clasificadas | %Hit Frase |
|---|---|---|---|---|---|---|---|---|
| EmoLex_Emo | 53 | 0,58 | 14,96 % | 3,62 | 2 | 8 | 92 | 57,61 % |
| RedPal-VF | 0 | 0,00 | 12,60 % | 2,00 | 2 | 2 | 4 | 0,00 % |

Tabla 5-5 Recall Ambiguo para Redpal-VF y EmoLEX español.

Para el caso analizado en la Tabla 5-5, EmoLEX analizo 92 frases y Redpal-VF 4 frases, donde EmoLEX tiene un Recall 58% y Redpal-VF un *Recall* de 0%. El *Recall* de EmoLEX se considera aceptable, además podemos decir que mientras EmoLEX hace hit (*tp*) a un 57,61% de las frases analizadas RedPal-VF hit (*tp*) a un 0% de las frases analizadas, las coberturas promedio son 14,96% para EmoLEX y 12,60% para RedPal-VF. Podemos destacar que la versión final de RedPal-VF redujo en forma significativa la ambigüedad en el análisis con respecto a la versión inicial, de 56 frase a solo 4 frases, aunque no pudo identificar una clase relevante en esas 4 frases. EmoLEX tiene un promedio de 3,62 al identificar la clase

relevante y una variación de 2 a 8 clases relevantes identificadas, mientras que RedPal tienen un promedio de 2 y una variación de 2 a 2.

Análisis combinado para la versión final de RedPal que llamaremos RedPal-VF será *Recall Hit* y *Recall Ambiguo*, esta versión contiene todo el proceso que se definieron para la creación del recurso léxico, desambiguaciones, normalizaciones y creación de la red, comparada con EmoLEX en español y una cobertura de 1% a 100% para ambos lexicones, los resultados en la Tabla 5-6.

| Lexicón | Hit Total | Recall Total | Cobertura | Prom. Clase Relevante | Mín. Clase Relevante | Máx. Clase Relevante | Frases Clasificadas | %Hit Frase |
|---|---|---|---|---|---|---|---|---|
| EmoLex_Emo | 62 | 0,49 | 15,57 % | 2,91 | 1 | 8 | 126 | 49,21 % |
| RedPal-VF | 69 | 0,21 | 15,29 % | 1,01 | 1 | 2 | 331 | 20,85 % |

Tabla 5-6 Recall Hit y Ambiguo para Redpal-VF y EmoLEX español

Para el análisis combinado de la Tabla 5-6, EmoLEX analizo 126 frases y Redpal-VF 331 frases, donde EmoLEX tiene un Recall 49% y Redpal-VF un *Recall* de 21%. El *Recall* de EmoLEX se considera aceptable, para RedPal-VF el *Recall* se considera deficiente, además podemos indicar que EmoLEX hace hit (*tp*) a un 49,21% de las frases analizadas RedPal-VF hit (*tp*) a un 20,85% de las frases analizadas, las coberturas promedio son 15,57% para EmoLEX y 15,29% para RedPal-VF. Se desataca que RedPal-VF hace un 10% más de hit totales *(tp)* que EmoLEX y analiza 166% más de frases. EmoLEX identifica en promedio 2,91 clases relevantes con una variación de uno a ocho clases relevantes, mientras RedPal-VF tiene un promedio de 1,01 clases relevantes una variación de uno a dos clases relevantes, RedPal-VF muestra una ambigüedad casi nula al identificar las clases relevantes.

### 5.1.7 RedPal en enfoque hibrido

Tal como fue mencionado antes, el recurso léxico es considerado uno de los elementos, importante, que debe ser utilizado en los análisis de emociones bajo un enfoque híbrido. Para dar cuenta de esto, se utilizará el Modelo de análisis para detectar agresividad propuesto en [60]. Para la evaluación se utilizó un corpus 1470 comentarios recopilados de *Twitter* en Chile, etiquetados en agresivos o no agresivos. El corpus balanceado se compone de 796 instancias.

Los modelos para la detección de agresividad que son comparados:

- Modelo 1: Sólo incorpora el análisis de ML base TF-IDF (no usa lexicones) sobre los comentarios etiquetados como agresivos/no agresivos

- Modelo 2: Incorpora el análisis de ML más el vector de intensidad de emociones
- Modelo 3: Incorpora el análisis de ML, el vector de intensidad de emociones y TF-IDF (frecuencias).

Para este trabajo sólo se utiliza el algoritmo *Support Vector machines*. En la tabla se resumen los resultados en la métrica *accuracy*.

| | ***Support Vector machines accuracy*** |
|---|---|
| **MODELO 1** | 0,774 |
| **MODELO 2** | 0.753 |
| **MODELO 3** | 0.803 |

Tabla 5-7 Resultados del análisis hibrido según modelo

Tal como se observa en la Tabla 5-7, la integración de RedPal en el análisis de emociones permite mejorar los resultados de los análisis basado sólo en *Machine Learning*.

## 5.2 Conclusiones del Capítulo

Las conclusiones de este capítulo se revisarán en extenso en el capítulo de discusión de los resultados de la evaluación.

# 6 Discusión de los resultados de la evaluación

A continuación, se presenta la discusión de los resultados de la evaluación de RedPal recurso léxico afectivo en español basado en la emoción expresada en cada palabra.

## 6.1 Resultados del análisis de emociones basado en lexicón RedPal

El resultado del experimento donde se compara RedPal en dos versiones con EmoLEX, una inicial donde no se integraban la desambiguación, normalización y creación de la red de emociones basado en intensidad, y otra donde si incluían estos aspectos, se resume en la Tabla 6-1. Un aspecto destacable de RedPal es que respecto a su versión inicial versus la final se redujo casi al mínimo la ambigüedad en la identificación de la clase afectiva relevante al evaluar una frase, la versión final obtiene el 100% de la clasificación identificando la clase relevante. Además de obtener números muy superiores a EmoLEX en este aspecto, el cual identifica una clase relevante solo en un 10% de los casos.

| Análisis | Lexicón | Hits | Recall | Cobertura | Prom. Clase Relevante | Frases Clasificadas | %Hit Frase |
|---|---|---|---|---|---|---|---|
| Recall Hit | EmoLex_Emo | 9 | 0,26 | 17,21 % | 1,00 | 34 | 26,47 % |
| Recall Hit | RedPal-V0 | 45 | 0,16 | 15,71 % | 1,00 | 275 | 16,36 % |
| Recall Hit | RedPal-VF | 69 | 0,21 | 15,32 % | 1,00 | 327 | 21,10 % |
| Recall Ambiguo | EmoLex_Emo | 53 | 0,58 | 14,96 % | 3,62 | 92 | 57,61 % |
| Recall Ambiguo | RedPal-V0 | 38 | 0,68 | 13,23 % | 2,16 | 56 | 67,86 % |
| Recall Ambiguo | RedPal-VF | 0 | 0,00 | 12,60 % | 2,00 | 4 | 0,00 % |
| Recall Hit y Ambiguo | EmoLex_Emo | 62 | 0,49 | 15,57 % | 2,91 | 126 | 49,21 % |
| Recall Hit y Ambiguo | RedPal-V0 | 83 | 0,25 | 15,29 % | 1,20 | 331 | 25,08 % |
| Recall Hit y Ambiguo | RedPal-VF | 69 | 0,21 | 15,29 % | 1,01 | 331 | 20,85 % |

Tabla 6-1 Resumen de evaluaciones.

Además, se presentan resultados de la clasificación de agresividad con un enfoque híbrido, en este caso utilizando *machine learning* junto con RedPal e información estadística del corpus.

Los resultados obtenidos, en comparación con otros trabajos similares de análisis de emociones, en términos generales, parecen poco atractivos, pero tomando en cuenta la casi nula disponibilidad de estos recursos etiquetados con intensidad en el idioma español, se podrían considerar promisorios. RedPal es un recurso léxico que no contiene todas las palabras de un idioma, sino que solo las que reflejan una emoción o afecto, por lo tanto, no es posible resultados destacables ya que solo contienen un subconjunto acotado del idioma.

Además, RedPal está pensado para complementar su análisis con enfoques híbridos como *machine learning.*

Otros aspectos hacen que RedPal aún tenga potencial por explotar, primero con un proceso definido para su expansión, actualización y retroalimentación constante. Luego la casi nula ambigüedad al realizar el análisis, en comparación a EmoLEX que sólo logra en un 10% de las frases analizadas sin ambigüedad. Se debe destacar que EmoLEX es un lexicón que, si bien este etiquetado en las ocho clases afectivas de Plutchik, todas las palabras la misma intensidad afectiva, es por esta la razón la alta ambigüedad, la cual casi se podría comparar a un análisis por densidad de palabras.

Otros aspectos de RedPal son la ya implementada vista de red o árbol, y su sistema de retroalimentación a través de encuestas, lo que permitirá que la red léxica se acerque a los parámetros aceptables teóricos para árboles con la cantidad de elementos que tiene RedPal. Además de contar con la base del recurso para la incorporación de palabras y regeneraciones, que permitirán el crecimiento organizado del recurso.

Además, el recurso léxico al incorporarlo a un modelo híbrido clasificación aumenta su efectividad y como lo mencionamos anteriormente es que RedPal pueda complementar distintos análisis, como vemos en la Figura 6-1.

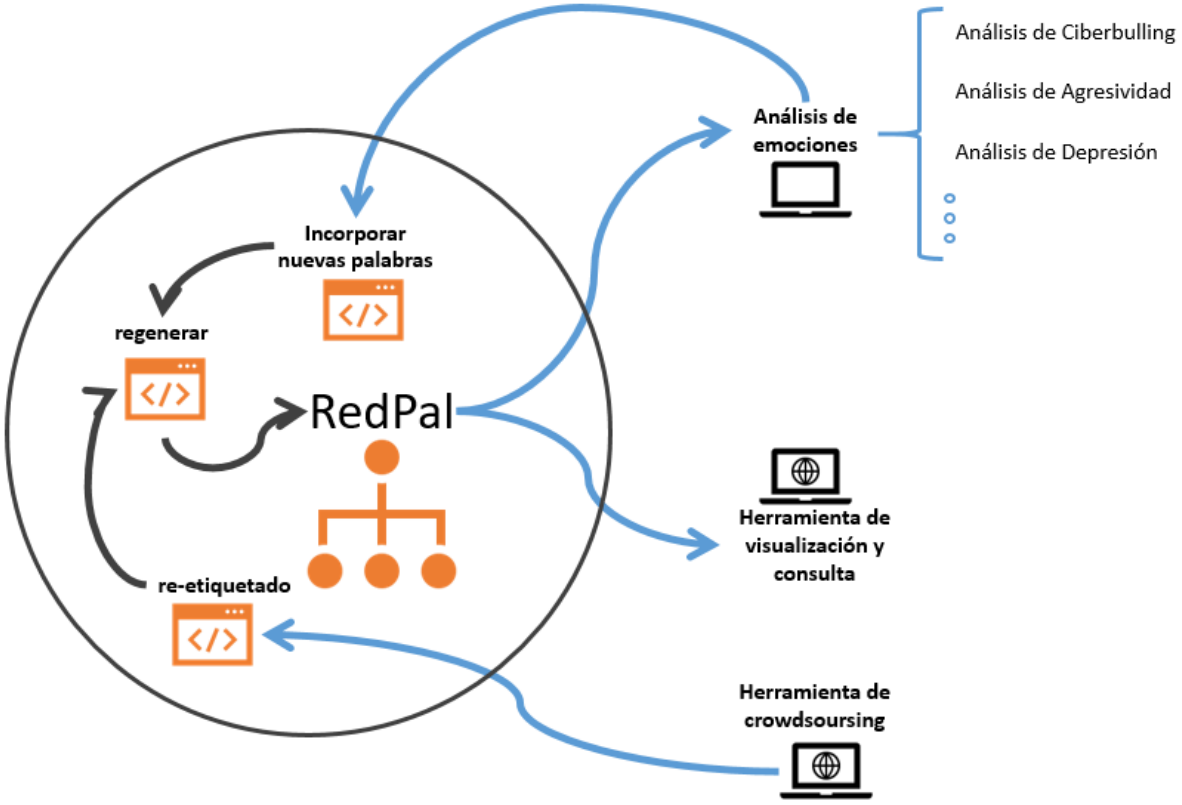



Figura 6-1 Proceso e interacción de RedPal

# 7 Trabajo futuro

En base a lo planteado en el capítulo anterior de discusión:

- Resulta necesario expandir RedPal, con nuevas palabras, frases coloquiales, emoticons que estén etiquetados en las ocho clases afectivas y que tengan o expresen una emoción, es decir que palabras que no expresen emoción no serán incorporadas a RedPal, el recurso léxico si bien define estas características no fueron evaluadas.
- Realizar un constante trabajo de retroalimentación a través de las encuestas para lograr la red léxica obtenida pueda llegar a parámetros teóricos aceptable.
- Definir nuevos criterios de incorporación automática de palabras.
- Evaluar las variaciones en los resultados del análisis de emociones utilizando las métricas implementadas en *Word-embedding* para corpus en español.

# 8 Conclusiones

En el capítulo final se presentas las conclusiones a partir de la realización de este trabajo, revisando el cumplimiento de la hipótesis y objetivos planteados, para finalizar con las conclusiones del trabajo

## 8.1 Conclusiones hipótesis y objetivos

En el capítulo 2 se presentó la hipótesis, objetivos generales y específicos del trabajo, el objetivo general era: “Definir un recurso léxico en español basado en la intensidad de la emoción expresada en cada palabra, que permita regenerar e incorporar nuevas palabras de forma automática, con el propósito de mejorar el análisis afectivo de textos”: El cual tenía los siguientes objetivos específicos que enumeraremos y mostraremos la forma en que se cumplieron.

- Analizar recursos léxico afectivo disponibles; información contenida, relaciones semánticas, estructura y formatos disponibles: Se efectúo una exhaustiva revisión sistemática de la literatura, en donde se reunieron los principales trabajos respecto a construcción de lexicones, análisis de sentimientos y análisis de sentimientos basados en lexicones. La revisión sistemática permitió contar con un marco teórico necesario para definir el estado del arte de las materias en estudio.
- Diseñar una estructura que soporte el recurso léxico que permita incorporar la intensidad de la emoción de cada palabra y considere criterios de transformación y regeneración para el enriquecimiento futuro del recurso: Se revisaron un gran número de trabajos relacionados con lexicones y se analizó ampliamente los distintos tipos de estructuras que los soportaban, definiendo una estructura jerárquica de árbol la cual esta soportaba en una base de datos relacional.
- Validar la efectividad de la propuesta a partir de un experimento que utilizando el recurso léxico propuesto mejore el rendimiento del análisis afectivo, incorporando un proceso de retroalimentación que regenere e incorpore nuevas palabras de forma automática: Se diseñaron 2 evaluaciones para el recurso léxico una basado en las características los cuales las diferencias de los lexicones conocidos en el análisis de emociones, las características relevantes que lo diferencian son el crecimiento, actualización constante, desambiguación, regeneración y retroalimentación. La segunda evaluación se realizó a partir de un experimento que consistía en comparar el

recurso lexicón con otro lexicón en español (EmoLEX), etiquetado en las ocho clases afectivas de Plutchik. Si bien en términos generales los números de efectividad son bajos o deficientes, el recurso léxico logra analizar una gran cantidad de frases del corpus utilizado, y reduciendo al mínimo la ambigüedad en la identificación de la clase relevante, incluso aumentado la efectividad al identificar esta clase relevante en comparación a la versión inicial del recurso. En comparación al recurso léxico inicial, los procesos de normalización, desambiguación y generación de la red léxica, la versión final de esta logro aumentar la efectividad de un *Recall* 16% a un 21%, cuando se identifica la clase relevante.

- Analizar objetivamente los resultados obtenidos y generar conclusiones y propuestas de trabajo futuro: Se realizó un análisis crítico y se presentaron distintas comparaciones de la evaluación del recurso léxico con versiones inicial y final. Los resultados presentan un aporte para líneas de trabajo futura relacionada con el enriquecimiento del lexicón y desambiguación, basados en intensidad afectiva y etiquetado en las ocho clases de Plutchik, además de seguir depurando el recurso léxico definido.

Con el cumplimiento de los objetivos específicos se cumplió el objetivo general propuesto, se definió un recurso léxico afectivo en español basado en intensidad de las palabras, el recurso definido contiene criterios para la regeneración e incorporación de nuevas palabras, y además mejora el análisis con respecto a una versión inicial que no contiene las características. Se pudo comprobar parcialmente la hipótesis definida en el capítulo 2 "***La representación de un recurso léxico afectivo en español basado en la intensidad emotiva y procesable computacionalmente, permite mejorar el rendimiento del clasificador en el análisis afectivo***", se definió un recurso el léxico en español basado en la intensidad de la emoción expresada y se le incorporaron características para la regeneración e incorporación de nuevas palabras, lo que produjo una mejora el análisis afectivo de textos en español, en comparación a la versión inicial.

## 8.2 Conclusiones generales

Este trabajo consistió en definir un recurso léxico afectivo en español basado en la intensidad de la emoción expresada por la palabra. Antes de definir el recurso léxico afectivo se realizó una validación y posterior expansión del lexicón inicial, esta tarea se llevó a cabo a partir de los *synset* de *WordNet* para extraer los sinónimos y su métrica *path* la cual fue tomada como intensidad afectiva inicial, además de realizar la traducción de estos sinónimos encontrados, e incorporados al lexicón. La revisión sistemática permitió identificar los aspectos más relevantes para diseñar el recurso léxico afectivo, además de su relación con el análisis de emociones basado en categorías definidas por Plutchik. Posterior a esto se comenzaron a definir las características necesarias para el recurso léxico afectivo en español, se definieron los criterios de desambiguación y normalización de las palabras y sus intensidades afectivas, eliminando las palabras que estaban en más de cuatro clases y además de normalizar el valor superior de las clases afectivas, tarea realizada se procedió a construir el árbol del recurso léxico llamado RedPal y la definición de los criterios para regeneración e incorporación de nuevas palabras, así también se defino la forma de retroalimentación del recurso a través de encuestas.

Para la evaluación se compararon las principales características de RedPal con los lexicones analizados en la revisión sistemática, y se enumeraron estas características en donde destaca RedPal por su crecimiento, actualización constante, desambiguación, regeneración y retroalimentación. Para el experimento de evaluación se comparó con otro lexicón en español etiquetado en las ocho categorías de Plutchik, EmoLEX, y aunque los resultados fueron de efectividad fueron deficientes un 21% mientras que EmoLEX obtiene un 49%, destaca la nula ambigüedad al identificar la clase relevante, RedPal 100% de su efectividad la obtiene identificando la clase relevante, mientras que EmoLEX sólo el 10% de su efectividad la obtiene identificando la clase relevante, por lo tanto RedPal resulta ser un recuso sin ambigüedad al identificar la clase relevante.

Con los resultados obtenidos se concluye que, el recurso léxico si cumple con lo planteado en el objetivo general y la hipótesis que era mejorar el rendimiento del análisis, lo cual se logra ya que al comparar el lexicón inicial si las características del recurso léxico definido su efectividad en el análisis aumenta de 16% a 21% cuando no existe ambigüedad en la identificación de clase afectiva relevante. Además, se concluye que el recurso léxico tiene mucho potencial basado en las características de crecimiento y expansión definidas para el recurso, lo que debe impactar positivamente en el análisis, clasificación y rendimiento.

# 9 Bibliografía


[1] W. Medhat, A. Hassan y H. Korashy, «Sentiment analysis algorithms and applications: A survey,» *Ain Shams Engineering Journal,* vol. 5, nº 4, pp. 1093-1113, 1 12 2014.

[2] H. Kaur, V. Mangat y Nidhi, «A survey of sentiment analysis techniques,» de *Proceedings of the International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2017*, 2017.

[3] S. K. Jain y P. Singh, «Systematic Survey on Sentiment Analysis,» de *ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications*, 2019.

[4] V. Pérez-Rosas, C. Banea y R. Mihalcea, «Learning Sentiment Lexicons in Spanish Identifying Visible Actions in Lifestyle Vlogs View project Personality, Images, and Text View project Learning Sentiment Lexicons in Spanish,» 2012.

[5] A. Shoukry y A. Rafea, «SATALex: Telecom Domain-specific Sentiment Lexicons for Egyptian and Gulf Arabic Dialects,» 2019.

[6] J. Redondo, I. Fraga, I. Padrón y M. Comesaña, «The Spanish adaptation of ANEW (Affective Norms for English Words),» *Behavior Research Methods,* vol. 39, nº 3, pp. 600-605, 8 2007.

[7] S. M. Mohammad y P. D. Turney, «CROWDSOURCING A WORD-EMOTION ASSOCIATION LEXICON,» *Computational Intelligence,* vol. 29, nº 3, pp. 436-465, 8 2013.

[8] J. Staiano y M. Guerini, «Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News,» de *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA, 2014.

[9] S. M. Mohammad y P. D. Turney, «Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon».

[10] S. M. Mohammad y P. D. Turney, «Crowdsourcing the Creation of a Word-Emotion Association Lexicon,» 2008.

[11] O. Araque, G. Zhu y C. A. Iglesias, «A semantic similarity-based perspective of affect lexicons for sentiment analysis,» *Knowledge-Based Systems,* vol. 165, pp. 346-359, 1 2 2019.

[12] A. Agirre, E. Laparra y G. Rigau, *Multilingual Central Repository version 3 . 0 : upgrading a very large lexical knowledge base,* 2011.

[13] C. Molina, A. Segura, C. Martinez, C. Vidal-Castro y C. Rubio-Manzano, «Improving the affective analysis in texts: Automatic method to detect affective intensity in lexicons based on Plutchik's wheel of emotions,» 2019.

[14] R. Plutchik, The emotions, University Press of America, 1991, p. 216.

[15] G. A. Miller y G. A., «WordNet: a lexical database for English,» *Communications of the ACM,* vol. 38, nº 11, pp. 39-41, 1 11 1995.

[16] M. Fares, A. Moufarrej, E. Jreij, J. Tekli y W. Grosky, «Unsupervised word-level affect analysis and propagation in a lexical knowledge graph,» *Knowledge-Based Systems,* vol. 165, pp. 432-459, 1 2 2019.

[17] R. Navigli y S. P. Ponzetto, «BabelNetXplorer,» de *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, New York, New York, USA, 2012.

[18] T. Tylenda, M. Sozio y G. Weikum, «Einstein,» de *Proceedings of the 20th international conference companion on World wide web - WWW '11*, New York, New York, USA, 2011.

[19] E. Agirre y A. Soroa, «Personalizing PageRank for word sense disambiguation,» de *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, Morristown, NJ, USA, 2009.

[20] B. Kitchenham y B. Kitchenham, «© Kitchenham, 2004 Procedures for Performing Systematic Reviews,» 2004.

[21] D. M. E. D. M. Hussein, «A survey on sentiment analysis challenges,» *Journal of King Saud University - Engineering Sciences,* vol. 30, nº 4, pp. 330-338, 1 10 2018.

[22] K. Ravi y V. Ravi, «A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,» *Knowledge-Based Systems,* vol. 89, pp. 14-46, 1 11 2015.

[23] K. Sailunaz, M. Dhaliwal, J. Rokne y R. Alhajj, «Emotion detection from text and speech: a survey,» *Social Network Analysis and Mining,* vol. 8, nº 1, 1 12 2018.

[24] A. Yadollahi, A. G. Shahraki y O. R. Zaiane, «Current state of text sentiment analysis from opinion to emotion mining,» *ACM Computing Surveys,* vol. 50, nº 2, 1 5 2017.

[25] M. V. Mäntylä, D. Graziotin y M. Kuutila, *The evolution of sentiment analysis—A review of research topics, venues, and top cited papers,* vol. 27, Elsevier Ireland Ltd, 2018, pp. 16-32.

[26] E. Cambria, A. Livingstone y A. Hussain, «The hourglass of emotions,» de *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[27] T. Solonchak y S. Pesina, «Lexicon Core and Its Functioning,» *Procedia - Social and Behavioral Sciences,* vol. 192, pp. 481-485, 6 2015.

[28] J. F. M. Burg, R. P. van de Riet y S. C. Chang, «A data dictionary as a Lexicon,» 1993.

[29] S. Rydin, «Building a hyponymy lexicon with hierarchical structure,» 2002.

[30] S. Pesina y L. G. Yusupova, «Words Functioning in Lexicon,» *Procedia - Social and Behavioral Sciences,* vol. 192, pp. 38-43, 6 2015.

[31] K. Allan, «Lexicon: Structure,» de *Encyclopedia of Language & Linguistics*, Elsevier, 2006, pp. 148-151.

[32] E.-J. Van Der Linden, «Incremental Processing and the Hierarchical Lexicon,» *Computational Linguistics,* vol. 18, nº 2, p. 219–238, 1992.

[33] G. P. Zarri, «A high-level representation language for the construction and use of large knowledge bases,» *Expert Systems With Applications,* vol. 1, nº 2, pp. 117-126, 1990.

[34] L. Li y B. R. Bryant, «An integrated parsing scheme for unification categorial grammar with object-oriented lexicon,» de *Proceedings of the ACM Symposium on Applied Computing*, 1994.

[35] A. Przepiórkowski, «A full-fledged hierarchical lexicon in LFG: the FrameNet approach,» *Bergen Language and Linguistics Studies,* vol. 8, nº 1, 23 11 2017.

[36] M. Hijzelendoorn y C. Cremers, «An Object-Oriented and Fast Lexicon for Semantic Generation,» 2010.

[37] J. Steinberger, P. Lenkova, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, R. Steinberger, H. Tanev, V. Zavarella y S. Vázquez, «Creating Sentiment Dictionaries via Triangulation,» de *2nd workshop on computational approaches to subjectivity and sentiment analysis. Association for Computational Linguistics*, 2011.

[38] S. Kiritchenko, X. Zhu y S. M. Mohammad, «Sentiment Analysis of Short Informal Texts,» 2014.

[39] F. L. Cruz, J. A. Troyano, B. Pontes y F. J. Ortega, «Building layered, multilingual sentiment lexicons at synset and lemma levels,» *Expert Systems with Applications,* vol. 41, nº 13, pp. 5984-5994, 1 10 2014.

[40] S. Baccianella, A. Esuli y F. Sebastiani, *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,* 2010.

[41] A. Esuli y F. Sebastiani, «SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining *,» 2007.

[42] A. Dey, M. Jenamani y J. J. Thakkar, «Senti-N-Gram: An n-gram lexicon for sentiment analysis,» *Expert Systems with Applications,* vol. 103, pp. 92-105, 1 8 2018.

[43] S. Mohammad, *#Emotional Tweets,* 2012, pp. 246-255.

[44] S. M. Mohammad y P. D. Turney, *Crowdsourcing the Creation of a Word – Emotion Association Lexicon,* 2010.

[45] S. M. Mohammad, «Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text,» de *Emotion Measurement*, Elsevier Inc., 2016, pp. 201-237.

[46] S. M. Mohammad, «Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words,» de *56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018.

[47] S. Poria, A. Gelbukh, E. Cambria, A. Hussain y G.-B. Huang, «EmoSenticSpace: A novel framework for affective common-sense reasoning,» 2014.

[48] J. F. Sánchez-Rada y C. A. Iglesias, «Onyx: A Linked Data approach to emotion representation,» *Information Processing and Management,* vol. 52, nº 1, pp. 99-114, 1 1 2016.

[49] S. M. Mohammad y S. Kiritchenko, «Using hashtags to capture fine emotion categories from tweets,» *Computational Intelligence,* vol. 31, nº 2, pp. 301-326, 1 5 2015.

[50] S. M. Mohammad, «Word Affect Intensities,» 27 4 2017.

[51] C. Strapparava y A. Valitutti, «WordNet-Affect: an Affective Extension of WordNet,» de *Proceedings of the 4th International Conference on Language Resources and Evaluation* , Lisbon, 2004.

[52] O. Araque, L. Gatti, J. Staiano y M. Guerini, «DepecheMood++: a Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques,» *IEEE Transactions on Affective Computing,* 8 10 2018.

[53] F. M. Plaza-del-Arco, M. T. Martín-Valdivia, L. A. Ureña-López y R. Mitkov, «Improved emotion recognition in Spanish social media through incorporation of lexical knowledge,» *Future Generation Computer Systems,* 9 2019.

[54] R. Rada, H. Mili, E. Bicknell y M. Blettner, «Development and Application of a Metric on Semantic Nets,» *IEEE Transactions on Systems, Man and Cybernetics,* vol. 19, nº 1, pp. 17-30, 1989.

[55] P. Resnik, «Using information content to evaluate semantic similarity in a taxonomy,» de *Proceeding IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*, Montreal, Quebec, Canada, 1995.

[56] J. A. Storer, An Introduction to Data Structures and Algorithms, Birkhäuser Boston, 2002.

[57] S. Anabalón y A. Segura, «Plataforma web de la red léxica afectiva en español, Tesis para optar al título de Ingeniería Civil en Informática[inédita],» 2020.

[58] T. Mikolov, K. Chen, G. Corrado y J. Dean, «Efficient estimation of word representations in vector space,» de *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.

[59] C. Martinez-Araneda, A. Segura, C. Vidal-Castro y J. Elgueta, «Is news really pessimistic? Sentiment Analysis ofChilean online newspaper headlines,» *Indian Journal of Science and Technology,* vol. 11, nº 22, pp. 1-8, 1 6 2018.

[60] M. Lepe, A. Segura y C. Vidal_Castro, «Modelos Híbridos basados en Lexicones y Machine Learning para la detección de agresividad sobre textos en el idioma Español, Tesis para optar al grado de Magister en Informática[inédita],» Universidad del Bío Bío, Chile, 2020.

[61] Y. Goldberg y O. Levy, «word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method,» 15 2 2014.

# 10 Anexos

## 10.1 Revisión sistemática de la literatura

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 1 | 2012 | #Emotional Tweets | Crea un corpus con los #hashtags de la emocion asociado al tweet | Inclusión |
| 2 | 1993 | A data dictionary as a Lexicon | En lingüística un lexicón contiene información sintáctica y semántica.<br>Existen distintos tipos de lexicón:<br>Lexicones Gramáticos<br>Lexicones de Conceptos(comunes y específicos)<br>Lexicones de taxonomía(is_a) | Inclusión |
| 3 | 2017 | A full-fledged hierarchical lexicon in LFG: the FrameNet approach | Teorías lingüísticas basadas en restricciones y proyectos lexicográficos, la información se organiza jerárquicamente. Jerarquía basada en *Valencia* | Inclusión |
| 4 | 1990 | A high-level representation language for the construction and use of large knowledge bases | Representación de la gramática y la representación del lenguaje | Inclusión |
| 5 | 2019 | A library for automatic natural language generation of spanish texts | Diccionario | Exclusión |
| 6 | 1992 | A reusable lexical database tool for machine translation | Diccionario/traductor | Exclusión |
| 7 | 2014 | A rule-based translation from written Spanish to Spanish Sign Language glosses | Diccionario/traductor | Exclusión |
| 8 | 2019 | A semantic similarity-based perspective of affect lexicons for sentiment analysis | Evalúan un nuevo método para análisis con lexicón: distancia jerárquica v/s la típica coincidencia de palabras. | Inclusión |
| 9 | 2017 | A survey of sentiment analysis techniques | Subjetivo: Representa un sentimiento<br>Objetivo: Contiene información que se puede comprobar<br>Polaridad: Positivo, negativo, neutral | Inclusión |
| 10 | 2015 | A survey on opinion mining and sentiment analysis: Tasks, approaches and applications | No relevante | Exclusión |
| 11 | 2018 | A survey on sentiment analysis challenges | No relevante | Exclusión |

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 12 | 1994 | An integrated parsing scheme for unification categorial grammar with object-oriented lexicon | Jerarquía de herencia para organizar un lexicón de gran tamaño y complejo, de manera eficiente.<br>Lexicón gramatical Implementado con OODB, base de datos orientada a objeto.<br>La red generada permite representar la semántica de cada palabra de manera uniforme en nodos y enlaces. | Inclusión |
| 13 | 2010 | An Object-Oriented and Fast Lexicon for Semantic Generation | Unificación de estructuras léxicas.<br>HPSG-Style(Head-driven phrase structure grammar, Gramática Sintagmática Nuclear) no relaciona estructuras de superficie con estructuras profundas.<br>DAG(Direct acylic graph o grafico acilico dirigido) estructura recursiva. | Inclusión |
| 14 | 2013 | An Optimized Formulation of Decision Tree Classifier | No corresponde al tema | Exclusión |
| 15 | 2012 | BabelNetXplorer | Los recurso como EuroWordNet, MCR,MENTA solo están disponibles para consulta, en inglés y sus códigos no están disponibles para desarrolladores.<br>La información de BabelNet está representada en un grafo dirigido donde los nodos representan conceptos y nombre de entidades, y los enlaces representan la relación entre ellos. | Inclusión |
| 16 | 2002 | Building a bilingual WordNet-like lexicon | Traductor | Exclusión |
| 17 | 2002 | Building a hyponymy lexicon with hierarchical structure | El lexicón es usado para apartar el conocimiento semántico en diferentes sistemas.<br>La evaluación de jerarquías semánticas o lexicones siempre han sido grandes desafíos. | Inclusión |
| 18 | 2014 | Building layered, multilingual sentiment lexicons at synset and lemma levels | Lexicón de sentimientos es un recurso léxico que contiene información acerca de la implicancias emocionales de las palabras.<br>Usa un grafo semántico de WordNet y SentiWordNet.<br>Se recalcula la polaridad por capas con PageRank InverseFlow | Inclusión |

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 19 | 2013 | CROWDSOURCING A WORD-EMOTION ASSOCIATION LEXICON | Crowdsoursing: Resolver una tarea grande distribuyéndola para que un gran número de personas la resuelva o responda, generalmente tiene pago asociado.<br><br>Es un lexicón en archivo plano en donde se indica la asociación de cada palabra por lo cual está esta repetida una vez por cada emoción y para positivo o negativo, cada palabra aparece 10 veces en el archivo. | Inclusión |
| 20 | 2010 | Crowdsourcing the Creation of a Word – Emotion Association Lexicon | Crowdsoursing: Resolver una tarea grande distribuyéndola para que un gran número de personas la resuelva o responda, generalmente tiene pago asociado.<br>Es un lexicón en archivo plano en donde se indica la asociación de cada palabra por lo cual esta esta repetida una vez por cada emoción y para positivo o negativo, cada palabra aparece 10 veces en el archivo. | Inclusión |
| 21 | 2008 | Crowdsourcing the Creation of a Word-Emotion Association Lexicon | Crowdsoursing: Resolver una tarea grande distribuyéndola para que un gran número de personas la resuelva o responda, generalmente tiene pago asociado.<br>Es un lexicón en archivo plano en donde se indica la asociación de cada palabra por lo cual está esta repetida una vez por cada emoción y para positivo o negativo, cada palabra aparece 10 veces en el archivo. | Inclusión |

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 22 | 2017 | Current state of text sentiment analysis from opinion to emotion mining | Detección de subjetividad: Saber si un texto es subjetivo. Generalmente un texto subjetivo expresa una opinión personal. Clasificación de la polaridad de la opinión: Determinar si el texto expresa un opinión positiva, negativa o neutral. Detección de emoción: Tarea de detectar si un texto expresa una emoción o no. similar a detección de subjetividad. Clasificación de Emociones: Tarea de detectar en forma más precisa la o las emociones existentes en un texto, y clasificarlas en un subconjunto definido. | Inclusión |
| 23 | 2014 | Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News | Anotación afectiva automática por crowdsourcing de una red social de noticias. | Inclusión |
| 24 | 2014 | Developing and Maintaining a WordNet: Procedures and Tools | Desarrollar de una herramienta tipo *WordNet* requiere un gran número de profesionales y tiempo, tanto para el desarrollo como para la precisión. | Inclusión |
| 25 | 2011 | Einstein | Constructor de conocimiento desde distintas fuentes | Exclusión |
| 26 | 2018 | Emotion detection from text and speech: a survey | Desde el punto de vista psicológico, las emociones humanas pueden ser identificadas y agrupado según el tipo de emoción, emoción, intensidad, y muchos otros parámetros.<br>Los modelos de emoción son una forma estructurada de definir varias emociones humanos según algún puntajes, rangos o dimensiones.<br>Basado en diferentes teorías de la emoción, emoción existente los modelos se pueden dividir en dos clases: categórico y Dimensional. | Inclusión |
| 27 | 2010 | Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon | Anotaciones manuales por crowdsourcing en Mechanical Turk.<br>BSW. Mejor-peor escalamiento. Best-Wors Scaling. | Inclusión |

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 28 | 2017 | Exploring hierarchical linguistic structure for aspect-based sentiment analysis | Estructuras jerárquicas para representar lexicón | Inclusión |
| 29 | 1994 | HPSG lexicon without lexical rules | HPSG Head Driven phrase structure grammar.<br>Las restricciones de un modelo léxico pueden hacerse cargo de todas las operaciones asignadas a las reglas léxicas | Exclusión |
| 30 | 2019 | Improved emotion recognition in Spanish social media through incorporation of lexical knowledge | El uso de recursos específicos mejora el reconocimiento de emociones.<br>El uso de lexicones solo traducidos no mejora la clasificación.<br>El mapeo de emoción a palabra depende de las diferencias culturales.<br>Existe un necesidad de crear lexicones emocionales distintos al inglés. | Inclusión |
| 31 | 1992 | Incremental Processing and the Hierarchical Lexicon | Las estructuras de un léxico jerárquico no solo son de gran importancia para la representación redundante de la información léxica, si no que también puede contribuir a la eficiencia del procesamiento del lenguaje natural. | Inclusión |
| 32 | 1993 | Introduction to WordNet: An On-line Lexical Database | XML y Matriz léxica | Inclusión |
| 33 | 2013 | Is there a language of sentiment? An analysis of lexical resources for sentiment analysis | Evaluación de 4 recursos léxicos | Exclusión |
| 34 | 2012 | Learning Sentiment Lexicons in Spanish Identifying Visible Actions in Lifestyle Vlogs View project Personality, Images, and Text View project Learning Sentiment Lexicons in Spanish | Framework basado en WordNet mejora el análisis de sentimientos en otros idiomas, mejora aún más usando Machine Learning | Inclusión |
| 35 | 1996 | Lexical rules | Diccionario | Exclusión |
| 36 | 2015 | Lexicon Core and Its Functioning | Inicialmente un lexicón era una lista de morfemas en un lenguaje específico. Luego se incluyó un conjunto de reglas básicas que operan un lexicón.<br>En el proceso de integración se pueden encontrar 2 tipos de unidades que definen cuan extensa es la generalización, estas unidades son los atributos diferenciados y los componentes de generalizados. | Inclusión |

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 37 | 2006 | Lexicon: Structure | Un lexicón es un contenedor de expresiones del lenguaje cuyo significado no es determinable a partir de los significados, por lo tanto, un usuario debe memorizar la combinación del lenguaje en forma y significado. | Inclusión |
| 38 | 2017 | Linked Data Models for Sentiment and Emotion Analysis in Social Networks | No corresponde al tema | Exclusión |
| 39 | 2012 | Modelling selectional preferences in a lexical hierarchy | No corresponde al tema | Exclusión |
| 40 | 2011 | Multilingual Central Repository version 3 . 0 : upgrading a very large lexical knowledge base | MCR es una traducción de WordNet a distintos idiomas, que mantiene su estructura y funciones. | Inclusión |
| 41 | 2016 | Onyx: A Linked Data approach to emotion representation | Ontología | Inclusión |
| 42 | 2009 | Personalizing PageRank for word sense disambiguation | A partir de la ontología propuesta en [41] definen las bases para una nueva estructura de grafos y una medida para el apoyo de procesos de Word Sence Dasanbiguation | Inclusión |
| 43 | 2019 | SATALex: Telecom Domain-specific Sentiment Lexicons for Egyptian and Gulf Arabic Dialects | Recurso léxico por polaridad , no supervisado.<br>Usa un lexicón específico. Combina un lexicón a nivel de palabra y otro a nivel de frase. | Inclusión |
| 44 | 2014 | Sentiment analysis algorithms and applications: A survey | Sentiment Analisys generalmente es considerado al análisis por polaridad. Detección de emociones trata de encajar más de una emoción en un texto. Como por ejemplo las 8 categoría de Plutchik, esta tarea es realizada por ML o lexicón pero el uso de lexicón es más usado. | Inclusión |
| 45 | 2016 | Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text | Áreas del analisis de sentimientos | Inclusión |
| 46 | 2018 | Senti-N-Gram: An n-gram lexicon for sentiment analysis | Lexicón de ngrams de intensidad de sentimiento(polaridad). Unigramas y bigramas | Inclusión |
| 47 | 2019 | Social information discovery enhanced by sentiment analysis techniques | No relevante | Exclusión |
| 48 | 2017 | Social media sentiment analysis: lexicon versus machine learning | No relevante | Exclusión |

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 49 | 2018 | Speech and Language Processing | Definiciones:<br>Detección de emociones Valence, Arousal, Dominance | Inclusión |
| 50 | 2019 | Systematic Survey on Sentiment Analysis | Otros lenguajes | Inclusión |
| 51 | 2014 | TexEmo: Conveying Emotion from Text-The Study | Varios tipos de emoción en texto, área poco estudiada | Inclusión |
| 52 | 1991 | The emotions | Plutchik | Inclusión |
| 53 | 2018 | The evolution of sentiment analysis—A review of research topics, venues, and top cited papers | Crecimiento del análisis de sentimiento | Inclusión |
| 54 | 2012 | The hourglass of emotions | Plutchik | Inclusión |
| 55 | 2007 | The Spanish adaptation of ANEW (Affective Norms for English Words) | encuestas a estudiantes de sicologia | Inclusión |
| 56 | 2019 | Unsupervised word-level affect analysis and propagation in a lexical knowledge graph | Aprendizaje supervisado basado en corpus, estadísticamente se hace coincidir las palabras con patrones textuales. Usando un enfoque no supervisado y lexicón hace coincidir palabras con las semillas del lexicón de sentimientos evaluando la similar dad semántica y distancia de referencia<br>Granularidad de palabra, frase, sentencia, documento y aspecto | Inclusión |
| 57 | 2015 | Using hashtags to capture fine emotion categories from tweets | Extraen las emociones automáticamente desde los hashtags de twitter generando unigrams, asociando con las emociones. | Inclusión |
| 58 | 1995 | Using information content to evaluate semantic similarity in a taxonomy | Resnik | Inclusión |
| 59 | 2017 | Word affect intensities | Lexicon con intensidad usa la escala BSW best-worst scaling. 6000 palabras. BSW: a partir de la palabra y su frecuencia en el corpus | Inclusión |
| 60 | 1995 | WordNet: a lexical database for English | WordNet | Inclusión |
| 61 | 2004 | WordNet::Similarity - Measuring the relatedness of concepts | Las medidas de similitud cuantifican cuánto se parecen dos conceptos, en función de la información contenida en IS_a una jerarquía. | Inclusión |

| N° | Año | Titulo | Observación general | Decisión |
|---|---|---|---|---|
| 62 | 2019 | WordNet2Vec: Corpora agnostic word vectorization method | Vectorización de texto | Inclusión |
| 63 | 2015 | Words Functioning in Lexicon | Las palabras ayudan a combinar 2 tipos de conocimientos y 2 niveles de conciencia lo verbal y no verbal. Operacional: Hacer coincidir el conocimiento del hablante con el de su compañero, con el objetivo de transferir conocimiento. Una palabra representa un cuerpo para un concepto o grupo de conceptos que aporta cierta información | Inclusión |
| 64 | 2008 | Opinion mining and sentiment analysis | Libro más citado Lexicón | Inclusión |
| 65 | 2012 | Sentiment analysis and opinion mining | Lexicón<br>análisis de los sentimientos<br>detección de opiniones | Inclusión |
| 66 | 2007 | SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining | Anotación automática de todos los synsets de WordNet | Inclusión |
| 67 | 2010 | SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining | Anotación automática de todos los synsets de WordNet | Inclusión |
| 68 | 2004 | WordNet-Affect: an Affective Extension of WordNet | WordNet-Affect | Inclusión |
| 69 | 2018 | Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words | Anotaciones manuales por crowdsourcing. Best–Worst Scaling | Inclusión |
| 70 | 2018 | DepecheMood++: a Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques | Anotación afectiva automática por crowdsourcing de una red social de noticias. | Inclusión |
| 71 | 2014 | EmoSenticSpace: A novel framework for affective common-sense reasoning | Anotación automatica efectiva. | Inclusión |
| 72 | 2011 | Creating Sentiment Dictionaries via Triangulation | Utilizando la extracción automática de términos del diccionario de subjetividad. Triangulación y expansión. | Inclusión |
| 73 | 2014 | Sentiment Analysis of Short Informal Texts | Método automático, a partir de tweets con hashtags de palabras de sentimiento. Entradas separadas para contextos afirmativos y negados. | Inclusión |

## 10.2 Implementación

### 10.2.1 Clasificador

Para el entorno de desarrollo del clasificador y los principales métodos utilizados.

Características del hardware:

- Windows 10 pro
- Procesador Intel Core i5-7500 @3.40GHz
- 8 GB de memoria RAM
- 500GB SSD

La implementación del clasificador fue desarrollada en Python versión 3.7.4 y los principales módulos usados fueron:

- Pyodbc: Permite la conexión a la base de datos donde está alojado el recurso léxico.
- Unicodedata: Permite la codificación y decodificación en distintos juegos de caracteres, para homogenizar el análisis de texto.
- Nltk: herramienta de procesamiento de lenguaje natural, usada para realizar la tokenización, pasar a minúsculas la palabras y eliminación de stopwords.
- Pandas: Ofrece estructuras y operaciones para manejar tablas de datos.
- Numpy: Soporte para vectores, arreglos y matrices.
- File Handling: Permite lectura y escritura de archivos

El clasificador exporta el análisis del corpus o frase en un archivo JSON. Esta información fue procesada con Power Bi versión Desktop lo que permitió generar un analizar la data generara por el clasificador en forma homogénea y generar los análisis respectivos, obteniendo tablas y gráficos, los que fueron presentados en el capítulo de experimentación

### 10.2.2 Plataforma Web red Léxica

Para él desarrollo de esta plataforma se utilizó diversos softwares, tales como:

**Servidor Web:**

- Nombre: Hypertext Preprocessor.
- Abreviación: PHP

- Versión: 7.1.31

- Nombre: JavaScript
- Abreviación: JS
- Versión: 3.4.1

- Nombre: Vis Network
- Abreviación: VisJS
- Versión: 6.1.0

**Base de Datos:**

- Nombre: MySQL
- Abreviación: MySQL
- Versión: 15.1 Distribución 10.4.6-MariaDB

## 10.3 Comparando resultados utilizando Lexicón intensidad

| Número de instancias de cada corpus | |
|---|---|
| Nombre Corpus | Número de instancias |
| Corpus Completo | 1000 |
| Corpus Balanceado | 796 |
| Corpus mujer | 1470 |
| Corpus unión | 2470 |

Tabla 10-1 Instancias de cada corpus

### 10.3.1 Modelo 1: Base TF-IDF (para referencia, no usa lexicones)

| | Support Vector machines | Naive Bayes | Random Forest |
|---|---|---|---|
| Corpus Completo | 0,796 | 0,646 | 0,799 |
| Corpus Balanceado | 0,774 | 0,669 | 0,799 |
| Corpus mujer | 0,895 | 0,744 | 0,893 |
| Corpus unión | 0,895 | 0,714 | 0,891 |

Tabla 10-2 Rendimientos algoritmos *Machine Learning* Modelo *baseline* (*Accuracy*)

### 10.3.2 Modelo 2: Vector Lexicón

| | *Support Vector machines Accuracy* Lexicón Intensidad | *Naive Bayes Accuracy* Lexicón Intensidad | *Random Forest Accuracy* Lexicón Intensidad |
|---|---|---|---|
| Corpus Completo | 0.8166 | 0.786 | 0.802 |
| Corpus Balanceado | 0.753 | 0.694 | 0.812 |
| Corpus mujer | 0.866 | 0.804 | 0.863 |
| Corpus unión | 0.844 | 0.790 | 0.835 |

Tabla 10-3 Rendimientos algoritmos *Machine Learning* modelo *Vector Lexicón*

### 10.3.3 Modelo 3: Vector Lexicón + TF-IDF

| | *Support Vector machines Accuracy* Lexicón Intensidad | *Naive Bayes Accuracy* Lexicón Intensidad | *Random Forest Accuracy* Lexicón Intensidad |
|---|---|---|---|
| Corpus Completo | 0.820 | 0.616 | 0.835 |
| Corpus Balanceado | 0.803 | 0.698 | 0.843 |
| Corpus mujer | 0.882 | 0.696 | 0.890 |
| Corpus unión | 0.871 | 0.681 | 0.886 |

Tabla 10-4 Rendimientos algoritmos *Machine Learning* modelo *Vector Lexicón + TF-IDF*