



UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

INGENIERÍA EN SOFTWARE

**“ENRIQUECIMIENTO DEL MODELO BASADO EN
REGLAS VADER A TRAVÉS DE LEXICONES”**

TESIS

QUE PRESENTA

KEVIN MEJÍA GONZÁLEZ

DIRECTORA: DRA. YULIA NIKOLAEVNA LEDENEVA

TIANGUISTENCO, MÉX.

Julio 2018

Declaración de originalidad del trabajo escrito

Mediante esta carta hago constar que el trabajo de tesis presentado en este documento es original porque cita debidamente los contenidos utilizados como soporte a la investigación presentada, por lo que exonero a la Universidad Autónoma del Estado de México de cualquier problema de derechos de propiedad intelectual.

Kevin Mejía González

Agradecimientos

A mis padres por el apoyo incondicional durante toda mi trayectoria universitaria, por creer en mí y guiarme para lograr mis metas los amo mucho.

A mi hermano Ivan por su ayuda en esta etapa.

A mi familia y especialmente a mi abuelita Ernestina[†] quien ha sido mi motivación en todas las metas que he logrado cumplir.

A mi novia Andrea quien me ha apoyado a superarme en todos mis proyectos.

A la Dra. Yulia Nikolaevna Ledeneva por su apoyo, asesoría y paciencia durante la investigación. Además de haberme brindado la oportunidad de aceptarme como su asesorado.

A mis revisores, la Maestra Griselda Areli Matías Mendoza, al Maestro Rafael Cruz Reyes y al Maestro Gerardo Arturo Ávila Vilchis por sus conocimientos y apoyo durante la investigación realizada.

Resumen

Hoy en día podemos apreciar que las redes sociales han tenido un gran impacto a nivel mundial, permitiendo compartir opiniones y emociones sobre diversos eventos sociales. Es por ello que han surgido diversos métodos computacionales capaces de analizar las opiniones y emociones compartidas por los usuarios de las redes sociales. Un ejemplo es el análisis de sentimientos.

El análisis de sentimientos es una tarea contemplada dentro del procesamiento del lenguaje natural y el aprendizaje automático, que permite analizar las opiniones y emociones con el fin de conocer las principales necesidades de los usuarios. A través de los años se ha podido observar el desarrollo de diversos métodos computacionales de análisis de sentimientos. En esta tesis se analiza el método *VADER*.

El método *VADER* se basa en reglas que clasifica las emociones de las opiniones de Twitter. El método utiliza un listado de palabras llamado lexicón, con el fin de clasificar las emociones que expresan las opiniones de Twitter.

Existen diversos lexicones los cuales pueden enriquecer el método y obtener una mejor clasificación de emociones. Sin embargo, no se sabe cuál de los lexicones presentados en el estado de arte enriquezca mejor el método basado en reglas *VADER*. Se utiliza el corpus de la competencia internacional WASSA 2017 en el idioma inglés.

En esta tesis se obtienen resultados del método *VADER* para 7 lexicones y 21 combinaciones de estos lexicones. Se realizan los experimentos con y sin preprocesamiento.

Tabla de contenido

Declaración de originalidad del trabajo escrito	ii
Agradecimientos	iii
Resumen	iv
ÍNDICE DE FIGURAS	viii
ÍNDICE DE TABLAS	ix
CAPÍTULO 1. INTRODUCCIÓN	1
1.1. Antecedentes.....	3
1.1.1. Análisis de sentimientos basado en métodos supervisados (aprendizaje automático y lexicones)	3
1.1.1.1. <i>SentiWordNet</i>	4
1.1.1.2. <i>SenticNet</i>	4
1.1.1.3. <i>SentiStrenght</i>	5
1.1.1.4. <i>VADER</i>	5
1.1.2. Análisis de sentimientos basado en métodos no supervisados (lexicones)	6
1.1.2.1. <i>General Inquirer (GI)</i>	6
1.1.2.2. Hu-Liu.....	7
1.1.2.3. LIWC	8
1.1.2.4. ANEW	8
1.1.2.5. SO-CAL.....	9
1.2. Planteamiento del problema	9
1.3. Objetivos	10
1.4. Hipótesis	10
1.5. Justificación.....	11
1.6. Organización de la tesis	11
CAPÍTULO 2. MARCO TEÓRICO.....	13
2.1. Red Social.....	13
2.2. Twitter	14
2.3. Emociones	16
2.4. Clasificación de emociones	16
2.5. Procesamiento del Lenguaje Natural.....	17
2.6. Aprendizaje automático	17
2.7. Análisis de sentimientos	18

2.7.1. Método supervisado.....	18
2.7.2. Método no supervisado.....	19
2.7.2.1. Lexicón.....	20
2.7.2.1.1. Tipos de lexicón	20
2.7.2.1.2. Selección de palabras.....	22
2.7.2.1.3. Positivos	24
2.7.2.1.4. Negativos.....	24
2.7.2.1.5. Neutrales	25
2.7.2.1.6. Abreviaciones	25
2.7.2.1.7. Emoticones	26
2.7.2.1.8. Negaciones.....	27
2.7.2.1.9. Intensificadores y modificadores	28
CAPÍTULO 3. ESTADO DEL ARTE	29
3.1. Trabajos relacionados	29
3.1.1. <i>VADER: A Parsomionious Rule-based Model for Sentiment Analysis of Social Media Text</i> (Hutto, 2014)	29
3.1.1.1. Lexicones de orientación semántica basada en polaridades	30
3.1.1.2. Lexicones de intensidad basados en valencias	31
3.1.2. <i>Comparing and Combining Sentiment Analysis Methods</i> (Goncalves, 2013) ..	34
3.1.3. <i>A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts</i> (Musto, 2014).....	40
3.1.4. <i>Enhancing Lexicon-Based Review Classification by Merging and Revising Sentiment Dictionaries</i> (Cho, 2013)	43
3.1.5. <i>Lexicon-Based Methods for Sentiment Analysis</i> (Taboada, 2011)	46
3.2. Resumen del capítulo.....	48
CAPÍTULO 4. MARCO METODOLÓGICO	51
4.1. Metodología propuesta.....	51
4.1.1. Entrada	52
4.1.2. Preprocesamiento.....	52
4.1.3. Método basado en reglas <i>VADER</i>	53
4.1.3.1. <i>VADER Sentiment</i>	53
4.1.3.2. Lexicones	54
4.1.3.3. Corpus.....	55
4.1.4. Salida.....	55

4.1.4.1. Enriquecimiento del método basado en reglas <i>VADER</i>	55
4.1.4.2. <i>F-Measure</i>	55
CAPÍTULO 5. EXPERIMENTOS Y RESULTADOS.....	57
5.1. Experimentación sin preprocesamiento.....	57
5.1.1. Lexicones individuales.....	57
5.1.2. Lexicones combinados en pares.....	58
5.1.3. Lexicones combinados.....	60
5.2. Experimentación con preprocesamiento.....	61
5.2.1. Lexicones individuales.....	61
5.2.2. Lexicones combinados en pares.....	62
5.2.3. Lexicones combinados.....	64
CAPÍTULO 6. CONCLUSIONES Y TRABAJO FUTURO.....	66
6.1. Trabajo Futuro.....	67
REFERENCIAS.....	68
ANEXO A. Lexicones.....	77
A.1. Lexicón <i>VADER</i>	77
A.2. Lexicón SentiStrength.....	78
A.3. Lexicón <i>SentiWordNet</i>	78
A.4. Lexicón <i>SenticNet</i>	79
A.5. Lexicón Hu-Liu.....	79
A.6. Lexicón SO-Cal.....	80
A.7. Lexicón ANEW.....	80
A.8. Lexicones combinados.....	81
ANEXO B. Corpus.....	82
B.1. Corpus WASSA 2017 sin preprocesamiento.....	82
B.2. Corpus WASSA 2017 con preprocesamiento.....	82
ANEXO C. Lista de comandos utilizados y scripts.....	84
C.1. Comandos.....	84
C.2. Scripts.....	84
C.3. Script <i>F-Measure</i>	85

ÍNDICE DE FIGURAS

Figura 1. Interfaz AMT.	32
Figura 2. Comparación de resultados para opiniones de Twitter (Goncalves, 2013).	38
Figura 3. Resultados de la combinación de todos los lexicones (Goncalves, 2013).	39
Figura 4. Evaluación para el corpus de SemEval (Musto, 2014).	42
Figura 5. Evaluación para el corpus STS 2013 (Musto, 2014).	42
Figura 6. Fusión de los lexicones.	46
Figura 7. Etapas del método propuesto.	51
Figura 8. Interfaz de resultado de lexicones combinados.	61
Figura 9. Interfaz de resultado de lexicones combinados con preprocesamiento.	65

ÍNDICE DE TABLAS

Tabla 1. Categorías de las palabras contenidas en el lexicon de LIWC (Hutto, 2014).....	8
Tabla 2. Porcentaje de usuarios de Twitter por país en el año 2014 (Heffring, 14).	15
Tabla 3. Comparación de los tipos de lexicones (Hernández, 2016).....	22
Tabla 4. Se muestra la clasificación de palabras mediante hashtags en opiniones de Twitter (Kouloumpis, 2011).	25
Tabla 5. Abreviaciones utilizadas dentro de textos en redes sociales.	26
Tabla 6. Partículas de negación.....	28
Tabla 7. Resultados de VADER con respecto a otros lexicones (Hutto, 2014).....	34
Tabla 8. Clasificación de emociones como polaridades (Goncalves, 2013).	36
Tabla 9. Selección de eventos y palabras clave de Twitter (Goncalves, 2013).	37
Tabla 10. Conjunto de datos evaluados por humanos (Goncalves, 2013).....	37
Tabla 11. Resultados de los dos mejores lexicones con respecto a la evaluación de opiniones de Twitter (Goncalves, 2013).....	39
Tabla 12. Evaluación de la exhaustividad y precisión de los lexicones Cho (2013).....	45
Tabla 13. Evaluación porcentual de cada lexicon cargado en el método SO-Cal (Taboada, 2011).	48
Tabla 14. Mejores lexicones obtenidos.	49
Tabla 15. Total de palabras con respecto a la combinación de lexicones en pares.	54
Tabla 16. Clasificación de opiniones con respecto a su polaridad.....	56
Tabla 17. Resultados obtenidos por cada lexicon cargado en el método basado en reglas VADER.	58
Tabla 18. Resultados obtenidos con respecto a la precisión de la combinación de los lexicones.....	59
Tabla 19. Resultados obtenidos con respecto a la exhaustividad de la combinación de los lexicones.....	59
Tabla 20. Resultados obtenidos con respecto a la medida F de la combinación de los lexicones.....	60
Tabla 21. Resultados de los lexicones individuales realizando el preprocesamiento del corpus de opiniones.....	62
Tabla 22. Resultados obtenidos con respecto a la precisión.....	63
Tabla 23. Resultados obtenidos con respecto a la exhaustividad.	63
Tabla 24. Resultados de la combinación los lexicones con respecto al resultado de la medida F.....	64

CAPÍTULO 1. INTRODUCCIÓN

Actualmente las redes sociales como Twitter¹, Facebook² e Instagram³ se han convertido en una necesidad de interacción humana en todo el mundo, utilizadas por los usuarios para compartir sentimientos y emociones.

Las emociones son los pensamientos y sentimientos subjetivos de cada usuario con respecto a una situación social o personal. Liu (2012) clasifica las emociones en amor, alegría, sorpresa, enojo, tristeza y miedo. Con el objetivo de analizar las emociones de los usuarios, se ha tenido la necesidad de aplicar análisis de sentimientos en redes sociales.

Análisis de sentimientos se define como el área de la computación, la cual analiza las emociones de los usuarios acerca de distintas áreas como productos, servicios, compañías y organizaciones con el objetivo de conocer las principales necesidades de los usuarios (Beigi, 2009) (Cambria, 2017) (Singh, 2018).

Las emociones mencionadas anteriormente expresan una idea sobre situaciones de interés social que pudieran indicar o predecir eventos importantes, por ejemplo, huelgas, movimientos en el mercado de valores, predicción de ventas y elecciones presidenciales (Makrehchi, 2013) (Montesinos, 2014) (Sasank, 2016).

Otro ejemplo de la implementación del análisis de sentimientos fue en las elecciones de 2016 para la presidencia de Estados Unidos, donde la campaña de Donald J. Tump contrató a la empresa Cambridge Analítica⁴, que realizó un estudio sobre la personalidad de los usuarios de Facebook, recabando así datos de 50 millones de usuarios, equivalente a la cuarta parte de los votantes de la elección para la presidencia (Ríos, 2018).

Utilizando un algoritmo de análisis de personalidades, Cambridge Analítica pudo analizar la personalidad de los usuarios mediante el número de emociones que cada

¹ <https://twitter.com/>

² <https://web.facebook.com/>

³ <https://www.instagram.com/>

⁴ <https://cambridgeanalytica.org>

usuario tenía publicado en su perfil de Facebook logrando “conocer” a cada usuario, mostrando anuncios, comerciales o propuestas de campaña de acuerdo con la personalidad de los usuarios.

Cabe mencionar, que Facebook utiliza su propia clasificación de emociones expresada en los siguientes estados: “me gusta”, “me encanta”, “me divierte”, “me asombra”, “me entristece” y “me enfada”.

A nivel internacional algunas organizaciones realizan competencias con el objetivo de conocer cómo se desempeñan diferentes tareas de análisis de sentimientos para lograr a obtener una mejor evaluación de sentimientos o emociones. Algunas de las competencias importantes son:

- La organización *codalab* ha realizado el concurso *Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis* (WASSA)⁵ desde el año 2010, clasificando tweets en emociones tales como felicidad, ira, tristeza y miedo (Balahur, 2017).
- El concurso de *Sentiment Analysis* de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN)⁶ llamado TASS se realiza desde el año 2012 con el objetivo de analizar opiniones de Twitter en español, obteniendo la clasificación de polaridades de las opiniones siendo positivas, negativas o neutrales (Villena, 2017).
- La competencia *International Workshop on Semantic Evaluation* (SemEval)⁷ se ha realizado desde el año 2006 con el objetivo de analizar opiniones de Twitter, obteniendo la polaridad positiva, negativa o neutral de cada opinión (Apidianaki, 2018).

⁵ <https://wt-public.emm4u.eu/wassa2018/index.htm>

⁶ <http://www.sepln.org/workshops/tass/2017/>

⁷ <http://alt.qcri.org/semeval2018/index.php?id=tasks>

1.1. Antecedentes

A través de los años se han realizado distintos trabajos para el análisis de sentimientos de las opiniones de las personas, desde los primeros trabajos cognitivistas hasta el análisis de emociones en las redes sociales (Dubiau, 2013).

Para ello se ha realizado un extenso estudio de análisis de opiniones expresadas, ya sea de forma escrita o hablada, con el objetivo de saber cómo se expresa lingüísticamente la valoración, el aplauso, la crítica o la opinión general (Aguado, 2010).

De acuerdo con los primeros estudios realizados sobre análisis de sentimientos, existen dos aproximaciones con respecto al análisis de sentimientos: métodos supervisados y métodos no supervisados (Taboada, 2011).

1.1.1. Análisis de sentimientos basado en métodos supervisados (aprendizaje automático y lexicones)

Los métodos de análisis de sentimientos supervisados utilizan algoritmos de clasificación como n-bayes, máquinas de soporte vectorial, máxima entropía, árboles de decisión y redes neuronales artificiales, utilizando un conjunto de datos de entrenamiento para realizar una clasificación correcta de los sentimientos con respecto a la polaridad o sentimiento que reflejan (Collomb, 2010).

Los métodos de análisis de sentimientos supervisados utilizan lexicones. Un lexicón se puede definir como el conjunto o lista de palabras que tiene como objetivo obtener de dichas palabras una etiqueta que permita indicar el grado de orientación semántica ya sea ésta positiva o negativa (Liu, 2012).

1.1.1.1. *SentiWordNet*

*SentiWordNet*⁸ fue elaborado mediante una aproximación supervisada y uso del lexicón. *SentiWordNet* fue creado con el propósito de realizar minería de opiniones aplicado a distintos contextos sociales, políticos y de mercado (Esuli, 2010).

El lexicón *SentiWordNet* contiene un conjunto de 147,306 palabras etiquetadas con una valencia de 0 a 1. Las palabras del lexicón *SentiWordNet* están agrupadas en *synsets* (Esuli, 2010). Dichos *synsets* son la clasificación de palabras que pertenecen al conjunto de adjetivos, sustantivos y verbos, como ejemplo, si tenemos la opinión de un Tweet “*This is bad and terrible*” el conjunto *synsets* sería $s = \{bad, terrible\}$ obteniendo así la polaridad y valencia de la opinión de Twitter siendo esta positiva, negativa o neutral.

1.1.1.2. *SenticNet*

*SenticNet*⁹ es un método supervisado y lexicón, con el fin de analizar sentimientos en redes sociales basando la clasificación de opiniones en polaridades positivas y negativas (Cambria, 2010).

Su lexicón contiene un total de 14,244 palabras etiquetadas con una valencia de -1 a 1, basado en la aplicación del Procesamiento del Lenguaje Natural (PLN). El lexicón de *SenticNet* clasifica mediante “conceptos” las palabras clave encontradas en una opinión, un ejemplo sería “*Boring, it’s Monday Morning*”, donde *Boring* y *Monday Morning* serían los conceptos clave extraídos de dicha opinión, dando como valencia a la palabra “*Boring*” -0.383 (negativa) y “*Monday Morning*” con un valor de 0.228 (positiva) (Cambria, 2010).

⁸ <http://SentiWordNet.isti.cnr.it/>

⁹ <http://sentic.net/>

1.1.1.3. *SentiStrenght*

*SentiStrenght*¹⁰ es un método supervisado y utiliza lexicón que toma en cuenta algunas características léxicas para el análisis de sentimientos (emoticones) (Thelwall, 2013).

El lexicón de *SentiStrenght* contiene un total de 2,662 palabras con una valencia de -5 a 5, evaluando palabras con polaridades positivas, negativas, intensas y débiles, así como emoticones (Thelwall, 2013).

1.1.1.4. *VADER*

Con el objetivo de realizar un análisis de sentimientos con una evaluación mayor a la de un humano se han creado métodos y lexicones los cuales evalúan características léxicas como acrónimos, emoticones, abreviaciones e iniciales.

Actualmente existen métodos, que, además de tener un lexicón basado en dichas características léxicas y métodos supervisados, emplean reglas las cuales determinan la clasificación de la opinión mediante el número de palabras negativas o positivas que contiene la opinión. Un ejemplo de estos métodos es *VADER*¹¹.

Value Aware Dictionary sEntiment Reasoner (*VADER*, por sus siglas en inglés) es un método que ha sido construido especialmente para realizar análisis de sentimientos en redes sociales.

El lexicón de *VADER* contiene un total de 7,517 palabras incluidos emoticones, abreviaciones, acrónimos e iniciales etiquetados mediante una valencia de -4 a 4, el lexicón de *VADER* se obtuvo aplicando la metodología de *Machine Learning*,

¹⁰ <http://sentistrength.wlv.ac.uk/>

¹¹ <https://github.com/cjhutto/VADERSentiment>

Wisdom-of-the-Crowd (la sabiduría de los grupos) y el uso de lexicones (LIWC, ANEW, GI).

La clasificación de la polaridad (positiva, negativa o neutra) de una opinión del método *VADER* se realiza mediante los valores de cada palabra en el lexicon del método.

Se demuestra que el método *VADER* basado en reglas ha obtenido clasificaciones correctas en opiniones de redes sociales, fue desarrollado en el trabajo de (Hutto, 2014), donde se describe la elaboración de dicha herramienta y la eficacia del método con respecto a otros métodos y lexicones.

1.1.2. Análisis de sentimientos basado en métodos no supervisados (lexicones)

Los métodos basados en lexicones clasifican la polaridad o intensidad de las opiniones con respecto a un diccionario que tiene clasificadas las palabras mediante polaridades o intensidades (Collomb, 2013).

1.1.2.1. *General Inquirer (GI)*

Desde la década de los 60's se ha tenido la necesidad de estudiar la conducta de la sociedad ante una situación de interés, es por ello que en 1962 la universidad de Harvard crea el primer método no supervisado *General Inquirer*¹² (GI, por sus siglas en inglés) (Stone, 1963).

¹² <http://www.wjh.harvard.edu/~inquirer/>

GI fue desarrollado con el objetivo de realizar un análisis objetivo, sistemático y cuantitativo de un conjunto de textos con el objetivo de obtener características “claras” del texto analizado.

GI cuenta con un lexicón que contiene 11,000 palabras clasificadas en 183 categorías (entre ellas emociones) siendo actualizado hasta ahora para realizar análisis de sentimientos en redes sociales.

1.1.2.2. Hu-Liu

Hu-Liu¹³ es un método no supervisado el cual tiene el propósito de analizar productos de venta en la red, así como opiniones en redes sociales.

El lexicón de Hu-Liu está compuesto por 6,800 palabras, 2006 positivas y 4,783 negativas, construido mediante la base de datos *English lexical*, agrupando las palabras de su lexicón en positivas y negativas (Hu, 2004).

Hu-Liu implementa tres reglas para el análisis de sentimientos:

1. Identifica características importantes acerca de un producto o tema de interés social de los usuarios a través de las opiniones que expresan.
2. Dichas características son evaluadas de acuerdo con el lexicón generado, obteniendo así la polaridad.
3. Obtiene la polaridad de la opinión (Hu, 2004).

¹³ <https://www.cs.uic.edu/>

1.1.2.3. LIWC

*Linguistic Inquiry and Word Count*¹⁴ (LIWC) es un método no supervisado el cual tiene como principal objetivo analizar características emocionales, cognitivas y estructurales.

Conteniendo en su lexicón 4500 palabras organizadas en 76 categorías, incluyendo 905 palabras categorizadas como positivas y negativas.

En la tabla 1 se muestra un ejemplo de las categorías en LIWC.

Tabla 1. Categorías de las palabras contenidas en el lexicón de LIWC (Hutto, 2014).

Categoría de LIWC	Ejemplos	No. De palabras
Emoción positiva	<i>Love, nice, good, great</i>	406
Emoción negativa	<i>Hurt, ugly, sad, bad, worse</i>	499

El lexicón de LIWC fue validado por psicólogos, sociólogos y lingüistas haciendo que la evaluación de análisis de sentimientos sea fiable Tausczik (2010). En la actualidad LIWC es implementado para obtener el sentimiento político en opiniones de Twitter prediciendo la depresión en cada opinión.

1.1.2.4. ANEW

*Affective Norms for English Words*¹⁵ (ANEW, por sus siglas en inglés) es un método de sentimientos creado mediante el método no supervisado, el cual tiene un enfoque

¹⁴ <http://liwc.wpengine.com/>

¹⁵ <https://arxiv.org/abs/1103.2903>

más preciso sobre análisis de opiniones en redes sociales, tomando en cuenta acrónimos inmersos en opiniones de redes sociales como ejemplo *WTF, LOL*.

El lexicón de ANEW se estructura por 1034 palabras, clasificadas con valencias de 1 a 9, donde la valencia igual 5 es una palabra neutral, menor a 5 es una palabra negativa y mayor a 5 es una palabra positiva (Nielsen, 2011).

1.1.2.5. SO-CAL

*Semantic Orientation CALculator*¹⁶ (SO-CAL) fue creado bajo el esquema no supervisado, el cual retoma el análisis de sentimientos para redes sociales.

El lexicón SO-CAL contiene más de 6000 palabras clasificadas mediante un rango de valencia de -5 a 5, obteniendo una orientación semántica ya sea la polaridad, la intensidad de las palabras, frases o textos. El lexicón SO-CAL fue creado mediante diccionarios que contienen adjetivos, sustantivos, verbos, adverbios, intensificadores y negaciones (Taboada, 2011).

1.2. Planteamiento del problema

Anteriormente se explicó que actualmente existen métodos y lexicones los cuales fueron elaborados mediante métodos supervisados y no supervisados.

Teniendo en cuenta que existen diversos lexicones los cuales carecen de características léxicas fundamentales ya sean acrónimos, emoticones, abreviaciones e iniciales para el análisis de opiniones en Twitter, así como la importancia de comparar y conocer los métodos y lexicones que clasifiquen los mejores resultados correspondientes para cada opinión con respecto a la polaridad

¹⁶ <https://www.sfu.ca/~mtaboada/nserc-project.html>

y emociones que expresan cada opinión evaluada, de acuerdo con lo anterior el problema de esta tesis se plantea de la siguiente manera:

¿Cómo enriquecer el método basado en reglas (*VADER*) para un corpus de Twitter en idioma inglés?

1.3. Objetivos

Objetivo general:

Enriquecer el método basado en reglas (*VADER*) a través 7 lexicones en un corpus de Twitter en el idioma inglés.

Objetivos específicos:

- Analizar el método basado en reglas (*VADER*).
- Evaluar el desempeño del método basado en reglas (*VADER*) con 7 diferentes lexicones.
- Evaluar el desempeño del método basado en reglas (*VADER*) utilizando combinaciones de lexicones.
- Evaluar los resultados de cada opinión del corpus de Twitter.
- Realizar el preprocesamiento del corpus de Twitter.
- Evaluar los resultados con otros estudios realizados en el estado del arte.

1.4. Hipótesis

Si se aplican 7 lexicones en un corpus de Twitter en idioma inglés, se podrá enriquecer el método basado en reglas (*VADER*).

1.5. Justificación

Debido a la gran cantidad de lexicones que existen en distintos analizadores de sentimientos, se desea comprobar qué lexicón aplicado en el método basado en reglas (*VADER*) ayuda a enriquecer al método.

1.6. Organización de la tesis

En el capítulo 1, se abordaron los antecedentes relacionados a los métodos y lexicones elaborados a través de los años para realizar análisis de sentimientos, teniendo en cuenta que el análisis de sentimientos se ha aplicado para analizar opiniones de diversos eventos a nivel mundial en redes sociales.

Posteriormente se introducen los elementos que guiarán la presente tesis como lo es el planteamiento del problema, objetivos, hipótesis y justificación.

En el capítulo 2, se definen los conceptos necesarios que describen al proceso del análisis de sentimientos, como el procesamiento del lenguaje natural y el aprendizaje automático. Se desarrollan los conceptos de método supervisado y método no supervisado, debido a que los métodos y lexicones que se describen en el estado del arte se desarrollan mediante dichos métodos supervisados y no supervisados. Por último, se describe que es un lexicón y su estructura.

En el capítulo 3, se presentan los principales trabajos relacionados al análisis de sentimientos y la comparación de los mejores lexicones, realizando una descripción de los trabajos para conocer la elaboración del método basado en reglas *VADER* y el método necesario para enriquecerlo mediante los mejores lexicones obtenidos.

En el capítulo 4, se describe a detalle la metodología utilizada en la presente tesis, así como el desarrollo de cada una de las partes que conforman a la misma, posteriormente se realiza una descripción de cada una de las etapas del método que se propone.

En el capítulo 5, se describen todas las pruebas realizadas para enriquecer el método basado en reglas *VADER*, mostrando a detalle las pruebas realizadas mediante la evaluación de los mejores lexicones de forma individual, combinados en pares y todos combinados. Se muestran los resultados obtenidos y se compara el desempeño del método *VADER* con respecto al corpus de Tweets.

Finalmente, en el capítulo 6 se muestran los mejores lexicones que enriquecieron al método basado en reglas *VADER*. Se describe el trabajo futuro con respecto al desarrollo de un lexicon.

CAPÍTULO 2. MARCO TEÓRICO

En este capítulo se muestran los conceptos importantes tales como: red social, Twitter, análisis de sentimientos y técnicas de análisis de sentimientos, ya sea elaborados mediante los métodos supervisados y métodos no supervisados.

2.1. Red Social

Distintos autores definen a una red social como, cito textualmente:

- *“Estructuras compuestas por personas conectadas por uno o varios tipos de relaciones (de amistad, de parentesco, de trabajo, ideológicas) con intereses comunes”* (Islas, 2013).
- *“Se trata de un servicio Web que permite a individuos construir un perfil público o semipúblico dentro de un sistema compartido articulando una lista de usuarios del sistema con los que compartir una conexión navegando a través de la lista de conexiones propias y de las del resto de usuarios”* (Gallego, 2010).
- *“Las Redes Sociales no son más que la evolución de las tradicionales maneras de comunicación del ser humano, que han avanzado con el uso de nuevos canales y herramientas, y que se basan en la co-creación, conocimiento colectivo y confianza generalizada”* (Merodio, 2010).
- *“Son formas de interacción social, definida como un intercambio dinámico entre personas, grupos e instituciones. Un sistema abierto y en construcción permanente que involucra a conjuntos que se identifican en las mismas necesidades y problemáticas y que se organizan para potenciar sus recursos. La intervención en red es un intento reflexivo y organizador de esas interacciones e intercambios, donde el sujeto se funda a sí mismo diferenciándose de otros”* (Ramos, 2009).

- *“Una red social es un sitio en la red cuya finalidad es permitir a los usuarios relacionarse, comunicarse, compartir contenido y crear comunidades, o como una herramienta de democratización de la información que transforma a las personas en receptores y en productores de contenidos”* (Urueña, 2010).

Mencionados los conceptos anteriores se puede saber que una red social es un servicio web en donde usuarios se relacionan, comparten o comunican necesidades comunes, con la finalidad de interactuar de acuerdo con un tema de interés actual (Ramos, 2009) (Gallego, 2010) (Merodio, 2010) (Urueña, 2010) (Islas, 2013).

2.2. Twitter

Debido a la popularidad que ha tenido la red social Twitter con respecto a los acontecimientos mundiales, se ha convertido en el principal objeto de estudio para conocer las emociones que tiene los usuarios con respecto a dichos acontecimientos.

Twitter es un microblogging fundado en el año de 2006 teniendo como principal objetivo otorgar un servicio de mensajería social, compartiendo mensajes con respecto a un tema de interés actual. Dichos mensajes tienen una estructura de no más de 280 caracteres llamados Tweets.

Teniendo alrededor de 400 millones de visitas mensuales compartiendo opiniones en 17 distintos lenguajes (entre los más populares inglés y español) Twitter se ha convertido en una red social para compartir emociones en opiniones (O'Reilly, 2012). Algunos autores explican los beneficios de Twitter, cito textualmente:

- *“Twitter como una fuente de información han permitido el desarrollo de aplicaciones e investigaciones sobre temas internacionales que tienen un impacto en la sociedad actual”* (Kumar, 2013).

- *“Los servicios de microblogging como Twitter han proporcionado a investigadores sociales información sobre reacciones públicas, con el objetivo de compartir millones de mensajes acerca de opiniones y emociones sobre distintos temas y discusiones sobre problemas actuales siendo estos políticos o religiosos, así como productos y servicios”* (Jones, 2014) (Pak, 2015).

Un estudio realizado por la revista *sysomos* en el año 2014, menciona que al menos en un día existe un 85.3% de tweets publicados, donde Estados Unidos de América (USA, por sus siglas en inglés) tiene el mayor porcentaje de usuarios de Twitter (Heffring, 14).

En la tabla 2 se muestra el porcentaje de usuarios por día en Twitter.

Tabla 2. Porcentaje de usuarios de Twitter por país en el año 2014 (Heffring, 14).

País	Porcentaje de usuarios de Twitter
Estados Unidos	62.14
Reino Unido	7.78
Canadá	5.69
Australia	2.80
Brasil	2.00
Alemania	1.51
Holanda	1.28
Francia	0.90
India	0.87
Sudáfrica	0.85
Japón	0.71
Filipinas	0.64
Noruega	0.63

2.3. Emociones

En Twitter las opiniones de los usuarios pueden clasificarse dependiendo la emoción que desean mostrar sobre un acontecimiento actual. Las emociones pueden mostrar distintos puntos de vista que tiene un usuario con respecto a lo que está viviendo en su entorno actual.

Las emociones se pueden definir como:

- Agitaciones o estados de ánimo producidos por ideas, recuerdos, deseos y sentimientos, ayudando a las personas a reaccionar con rapidez ante acontecimientos sociales o personales, siendo emociones positivas o negativas (Pallarés, 2010) (Fridja, 2011) (Santiago, 2018).

Debido a que las emociones son un estado de ánimo de las personas, algunos autores han clasificado las emociones internas de un humano, con respecto a la reacción de una situación actual.

2.4. Clasificación de emociones

De acuerdo con Pallarés (2010) y Santiago (2018) las emociones se clasifican de la siguiente forma:

- Miedo. Anticipación de una amenaza o peligro.
- Sorpresa. Permite una aproximación cognitiva de lo que está ocurriendo.
- Aversión o asco. Disgusto hacia aquello que tenemos en frente.
- Ira. Sentimiento de enojo ante una situación desagradable o amenazante.
- Alegría. Sensación de bienestar y de seguridad que sentimos cuando conseguimos algún deseo o meta.
- Tristeza. Pena, soledad, pesimismo ante la pérdida de algo importante.

En las opiniones de Twitter se ha aplicado esta clasificación de emociones, con el objetivo de conocer las necesidades de los usuarios

2.5. Procesamiento del Lenguaje Natural

El procesamiento del lenguaje natural (PLN) es una disciplina de la inteligencia artificial relacionada directamente con la lingüística, que tiene por objetivo lograr la comunicación entre los seres humanos y las computadoras, obteniendo la idea principal de un texto, documento u opinión, con la posibilidad de desarrollar sistemas capaces de realizar tareas de acuerdo con el lenguaje (Gelbukh, 2002) (Cortez, 2002) (Alberich, 2007).

Dubiau (2013) menciona que el procesamiento del lenguaje natural puede aplicarse al análisis de sentimientos con el objetivo de clasificar documentos, textos u opiniones a partir de la identificación y extracción de información subjetiva (opiniones, sentimientos o emociones).

2.6. Aprendizaje automático

El aprendizaje automático es el área de la inteligencia artificial el cual permite la detección automática de patrones en los datos con el objetivo de obtener conocimiento, permitiendo desarrollar sistemas capaces de aprender y obtener conocimiento por sí mismos, aplicando diversos algoritmos como máquinas de soporte vectorial, n-bayes y máxima entropía. Estos sistemas han tenido diversos usos como es el caso de la detección de fraudes en tarjetas de crédito, sistemas de filtrado de información y vehículos autónomos (Michel, 1997) (Nilsson, 1998) (Shaley, 2014) (Conway, 2012) (Hurwitz, 2018).

El aprendizaje automático se divide en dos categorías:

- Métodos supervisados
- Métodos no supervisados.

2.7. Análisis de sentimientos

Con el objetivo de realizar la clasificación de emociones en opiniones de Twitter, se explicarán las técnicas empleadas por el análisis de sentimientos, definiendo en primera instancia que es el análisis de sentimientos.

El análisis de sentimientos o minería de opiniones es una rama de la computación que permite analizar opiniones, sentimientos y emociones a ciertas áreas de interés social como productos, servicios, organizaciones, compañías, eventos y temas de interés actual, el cual se realiza mediante la aplicación del Procesamiento del Lenguaje Natural (PLN), Aprendizaje Automático e Inteligencia Artificial (IA), siendo que, dichas opiniones, pueden ser expresadas mediante Tweets, clasificadas en emociones o polaridades (Liu, 2012) (Beigi, 09) (Cambria, 2017) (Singh, 2018).

De acuerdo con la definición anteriormente desarrollada, se puede observar que el análisis de sentimientos se desarrolla mediante aplicación de métodos como PLN, Aprendizaje Automático, e IA entre otros, dichos métodos son llamados métodos supervisados. Uno de los puntos importantes que los autores mencionados no toman en cuenta para definir análisis de sentimientos es la aplicación de métodos no supervisados basándose principalmente de lexicones, los cuales serán definidos más adelante en este capítulo.

2.7.1. Método supervisado

El método supervisado se basa principalmente en dos conjuntos de datos, de entrenamiento y de pruebas, el conjunto de datos de entrenamiento tiene como propósito entrenar a los algoritmos supervisados entre ellos: Máquinas de Soporte

Vectorial (SVM), Naive Bayes (NB), Máxima Entropía (Singh 2018) entre otros, obteniendo así resultados correctos (Marsland 2015), encontrando patrones y relaciones en los datos. Una vez entrenados los algoritmos supervisados, se evalúan mediante el conjunto de datos de prueba para verificar el rendimiento que obtuvo cada algoritmo supervisado (Contreras, 2016).

Gutiérrez (2015) menciona la aplicación de método supervisado para el análisis de sentimientos, donde se tiene un conjunto de datos, que, para el caso de Twitter son las opiniones, y las respectivas clases clasificadas (emociones).

Con el propósito de clasificar correctamente las clases con respecto a su opinión, se debe realizar una comparación con un conjunto de datos de entrenamiento.

2.7.2. Método no supervisado

Este método se utiliza principalmente para evaluar una gran cantidad de datos no etiquetados, es decir, no trabaja con un conjunto de datos de entrenamiento. El método no supervisado trabaja bajo un esquema de clasificación manual para el análisis de documentos, textos u opiniones con el objetivo de agrupar los datos con respecto a dicha clasificación. Dicho método se utiliza principalmente para resolver problemas relacionados con el análisis de redes sociales, teniendo como principal conjunto de datos no etiquetados los tweets. Es por ello que para el análisis de sentimientos el método no supervisado trabaja con palabras o lexicones etiquetados previamente. Comparando la entrada de los datos (para este caso los Tweets) con las palabras o lexicones etiquetados con un valor previo, con el objetivo de obtener la clasificación de los datos de entrada (Mohri, 2012) (Marsland, 2015) (Contreras, 2016) (Hurwitz, 2018).

2.7.2.1. Lexicón

En la definición del método no supervisado, se mencionaba que para el análisis de sentimientos la evaluación de documentos, textos u opiniones se realiza mediante palabras o en su defecto un conjunto de palabras etiquetadas con un valor, este conjunto de palabras tiene por nombre Lexicón.

Un lexicón se puede definir como un conjunto de palabras o lemas finitas que contiene una orientación semántica (valencia o polaridad) del conjunto de palabras, siendo estas positivas o negativas. Con el objetivo de etiquetar un documento, texto, oración u opinión de acuerdo con su orientación semántica o polaridad (Zhang, 2011) (Liu, 2012) (Barbosa, 2015) (Hernández, 2016) (Becerra, 2017).

Para entender la construcción de un lexicón Liu (2012) menciona que la lista de palabras que componen un lexicón (palabras de sentimiento, palabras de opinión, palabras de polaridad o palabras de soporte de opinión), deben ser positivas y negativas. Las palabras positivas describen estados o cualidades deseadas en un conjunto de datos mientras que las palabras negativas muestran un estado no deseado.

Las palabras en un lexicón pueden ser divididas dentro de dos categorías, tipo base y tipo comparativo. Las palabras de tipo comparativo se utilizan para expresar opiniones comparativas y superlativas, como ejemplo tenemos palabras tales como “mejor”, “peor”, “peor”, “vencer”, etc. Mientras que las palabras de tipos base son las clasificadas mediante una polaridad. Para la construcción de un lexicón que permita un análisis de sentimientos se deben tomar en cuenta palabras de tipo base.

2.7.2.1.1. Tipos de lexicón

Con el objetivo de construir un lexicón basado en análisis de sentimientos, el lexicón se construye a partir de diversos tipos de lexicones:

- Lexicón manual: Es tipo de lexicón conlleva un trabajo exhaustivo, siendo necesario etiquetar de forma manual cada una de las palabras en el lexicón, dependiendo su orientación semántica o polaridad. Algunos autores como Hutto (2014) emplearon la ayuda “humana” para clasificar la polaridad de las palabras del lexicón creado (Hernández, 2016).
- Lexicón basado en un corpus lingüístico: Iniciando con una lista de un mínimo de palabras, este lexicón se basa principalmente en la agregación de palabras incluidas en documentos, textos u opiniones siendo estos de un gran tamaño. Dichas palabras se agregan al lexicón de acuerdo con la orientación semántica que poseen en el corpus (Liu, 2012) (Hernández, 2016).
- Lexicón basado en la aproximación de diccionarios: Se construye a partir de diccionarios establecidos (*WordNet* y *bootstrap*) los cuales contienen sinónimos y antónimos obteniendo de ellos la polaridad positiva o negativa (Liu, 2012) (Hernández, 2016). Goncalves (2013) explica que las palabras sinónimas se refieren a la agrupación de adjetivos, sustantivos y verbos evaluados de acuerdo con su orientación semántica o polaridad. Un ejemplo de ello serían la agrupación de sinónimos {malo, terrible, perverso}.

En la tabla 3 se muestra el tiempo de generación de un lexicón con respecto a la cantidad de palabras, la expresión del dominio de cada lexicón y el idioma en el cual se construye el lexicón.

Tabla 3. Comparación de los tipos de lexicones (Hernández, 2016).

Tipo de lexicón	Manual	Diccionario	Corpus
Tiempo de generación	Alto	Bajo	Medio
Cantidad de palabras del idioma	6000 a 9000	10,000 a 25,000	12,000 a 25,000
Expresiones propias del dominio	Media	Baja	Alta
Características propias de la región	Media	Baja	Alta

2.7.2.1.2. Selección de palabras

En la descripción de los tipos de lexicones, se menciona que debe realizarse una selección de palabras con respecto al conjunto de documentos, textos u opiniones. Es por ello que la selección de palabras para un lexicón se realiza mediante el número de veces que una palabra aparece escrita en un documento, texto u opinión sea esta positiva o negativa para obtener que palabras tienen una mayor frecuencia dentro de del conjunto de datos (Montesinos, 2014).

Con el objetivo de obtener la polaridad de cada palabra en el lexicón Hernández (2016) demuestra una aproximación algebraica, utilizando diversas fórmulas para los cálculos.

$$\text{Polaridad}(p) = \text{PMI}(p, \text{positivo}) - \text{PMI}(p, \text{negativo}) \quad (1)$$

La polaridad de las palabras es la resta del punto de información mutua (*PointWise Mutual Information PMI*¹⁷, por sus siglas en inglés) de la frecuencia de palabras positivas menos el punto de información mutua de la frecuencia de las palabras negativas. El punto de información mutua se refiere a la medida de asociación de la frecuencia de palabras siendo estas positivas o negativas.

Para obtener el *PMI* de las palabras positivas o negativas, se calcula mediante la siguiente fórmula.

$$\text{PMI}(p, \text{positivo}) = \log_2 \frac{\text{Frec}(w, \text{positivo}) * N}{\text{Frec}(w) * \text{Frec}(\text{positivo})} \quad (2)$$

Dónde:

$\text{Frec}(p, \text{positivo})$ = es la frecuencia de palabras positivas en un conjunto de datos, considerando un ajuste, dicho ajuste se refiere al cálculo de polaridad evitando la división entre 0, esto se da solamente cuando no existe una palabra en ambos conjuntos de datos positivos y negativos.

Calculando como la frecuencia de las palabras p positivas más el ajuste r .

$$\text{Frec}(p, \text{positivo}) = \text{frecuencia}_p + r \quad (3)$$

$\text{Frec}(p)$: frecuencia total de p en el conjunto de datos. Se calcula mediante la suma de $\text{Frec}(p, \text{positivo})$ y $\text{Frec}(p, \text{negativo})$

$$\text{Frec}(p) = \text{Frec}(p, \text{positivo}) + \text{Frec}(p, \text{negativo}) \quad (4)$$

$\text{Frec}(\text{positivo})$: Número total de palabras dentro del lexicón.

N : Sumatoria de la frecuencia de las palabras en el conjunto de datos.

¹⁷ https://es.wikipedia.org/wiki/Punto_de_informaci%C3%B3n_mutua

$$N = \sum_{p \text{ en lexicón}} Frec(p) \quad (5)$$

2.7.2.1.3. Positivos

Una vez explicado cómo se seleccionan las palabras dependiendo su orientación semántica o polaridad, se definirán que son las palabras positivas, negativas y neutrales.

De acuerdo con la real academia española la palabra se define como:

“f. Inform. Unidad lingüística, dotada generalmente de significado, que se separa de las demás mediante pausas potenciales en la pronunciación y blancos en la escritura.”.

De acuerdo con la real academia española el adjetivo positivo se define como.

“f. Adj. Afirmativo o que expresa afirmación o aceptación.”.

De acuerdo con lo anterior las palabras positivas se definen como: La unidad lingüística generalmente con un significado afirmativo o que expresa afirmación.

2.7.2.1.4. Negativos

De acuerdo con la real academia española el adjetivo negativo se define como.

“f. Adj. Que produce algún daño o perjuicio o resulta desfavorable para algo.”

Las palabras negativas se definen cómo: La unidad lingüística generalmente con un significado que produce algún daño o perjuicio, siendo desfavorable para algo.

2.7.2.1.5. Neutrales

De acuerdo con la real academia española el adjetivo neutral se define como.

“f. Adj. Que no se inclina en favor de ninguna de las partes opuestas o enfrentadas.”

Las palabras neutras se definen cómo: La unidad lingüística generalmente con un significado que no se inclina en favor a ninguna de las partes opuestas.

En la tabla 4 se muestra un ejemplo de las palabras seleccionadas de acuerdo con su orientación semántica ya sean positivas, negativas o neutrales.

Tabla 4. Se muestra la clasificación de palabras mediante hashtags en opiniones de Twittter (Kouloumpis, 2011).

Positivos	<i>#iloveitwhen, #thingsilike, #bestfeeling, #besfeelingever, #omgthatssotruer, #imthankfulfor, #thingslove, #success</i>
Negativos	<i>#fail, #epicfail, #nevertrust, #worst, #worse, #itsnotokay</i>
Neutrales	<i>#job, #tweetajob, #hiring, #cnn</i>

Es importante conocer las palabras positivas, negativas y neutras al realizar análisis de sentimientos ya que se puede conocer las necesidades o quejas que tienen los usuarios de las redes sociales es decir si tienen una opinión favorable o crítica ante un suceso mundial o de ámbitos comerciales.

2.7.2.1.6. Abreviaciones

Las abreviaciones dentro de un lexicón son de gran utilidad al momento de realizar análisis de sentimientos, evaluando la polaridad u orientación semántica de cada abreviación que se tenga en el lexicón, con el propósito de obtener una mejor clasificación de sentimientos y emociones en los conjuntos de datos a analizar.

Teniendo en cuenta que Twitter tiene como un límite de 280 caracteres al escribir una opinión, se ha tenido la necesidad de lograr la practicidad en la comunicación entre usuarios de Twitter por simplificar o abreviar las palabras escritas en una opinión (Eres, 2009).

El uso de estas abreviaciones dentro de las redes sociales ha crecido de forma exponencial dentro de textos en línea, teniendo la necesidad de agregarlos a los lexicones para el análisis de sentimientos. Hutto (2014) elabora su lexicón para análisis de sentimientos agregando las abreviaciones más usadas dentro de textos de redes sociales.

En la tabla 5 se describen algunas abreviaciones utilizadas en redes sociales.

Tabla 5. Abreviaciones utilizadas dentro de textos en redes sociales.

Abreviatura	Significado
100mpre	Siempre
Aki	Aquí
LOL	<i>Laughing out loud</i> – Reír a carcajadas
DM	<i>Direct Message</i> – Mensaje privado
FB	<i>Facebook</i>
ILY	<i>I Love You</i> – Te amo
NVM	<i>Nevermind</i> - Olvídalo

2.7.2.1.7. Emoticones

Por otro lado, los emoticones han tenido un gran uso en textos de redes sociales, ya sea para comunicar de manera enfatizada una emoción o sentimiento positivo o negativo. Es por ello que se han agregado al momento de elaborar un lexicón, con el objetivo de obtener un mejor resultado al realizar un análisis de sentimientos.

Un emoticono (*emotion* que significa emoción y *icon* ícono) se puede representar como una serie de caracteres ASCII representando una cara humana expresando una emoción representados mediante signos de puntuación y paréntesis (Montesinos, 2014) (Amores, 2016).

Tomando un rol importante en las conversaciones en redes sociales los emoticonos se pueden clasificar de acuerdo con su escritura, dicha clasificación se da por emoticonos de código occidental y emoticonos de código oriental. Los emoticonos de código occidental generalmente están escritos mediante una orientación inclinada de 90° grados, un ejemplo de ellos son :-), :-), :-), :P, :S, 8-). Por otra parte, los emoticonos de código oriental se leen de forma vertical, un ejemplo de ellos sería (^o^), (^_^), (+_+), (@_@) (López, 2009).

2.7.2.1.8. Negaciones

Dentro de un lexicón existen ciertas partículas que no pueden ser incluidas, entre ellas las negaciones, no significa que las negaciones sean palabras negativas, las negaciones se encuentran incluidas en los documentos, textos u opiniones de redes sociales.

Desde las primeras investigaciones de minería de opiniones las negaciones han sido un problema al momento de analizar sentimientos en redes sociales, ya que se marcan como negadas todas aquellas opiniones (aun siendo positivas) que se encuentren después de una partícula negativa llegando a un signo de puntuación, conector o palabras gramaticales invirtiendo o reduciendo la polaridad u orientación semántica del documento, texto u opinión (Amores, 2016) (Farooq, 2016).

En la tabla 6 se muestran las partículas de negación utilizadas en redes sociales.

Tabla 6. Partículas de negación.

Clase Negación	Negaciones
Sintáctica	<i>no, not, rather, couldn't, wasn't, didn't, wouldn't, shouldn't weren't, don't, doesn't, haven't, hasn't, won't, wont, hadn't, never, none, nobody, nothing, neither, nor, nowhere, isn't, can't, cannot, mustn't, mightn't, shan't, without, needn't</i>
Diminutivos	<i>hardly, less, little, rarely, scarcely, seldom</i>

En algunos trabajos como en Hutto (2014) se han evaluados dichas partículas de manera empírica, ayudando de forma sustancial al momento de obtener una clasificación de emociones o polaridades de documentos, textos u opiniones en redes sociales.

2.7.2.1.9. Intensificadores y modificadores

Las palabras intensificadoras y modificadores son aquellas que pueden modificar, reducir o aumentar la polaridad o emoción de un documento, texto u opinión al realizar análisis de sentimientos. Los intensificadores pueden comportarse como reductores dependiendo del contexto del texto, haciendo énfasis en una palabra o frase. Los intensificadores o modificadores pueden dividirse de acuerdo con el contexto en el que se implementen, es decir, los adjetivos intensificadores modifican a los sustantivos, los adverbios de intensificación modifican a los verbos. Algunos de estos ejemplos pueden ser “muy”, “mas”, “bastante”, “poco”, “menos”, “casi” (Montesinos, 2014) (Amores, 2016).

De acuerdo con la investigación de Hutto (2014) se utilizan dichos modificadores o intensificadores dentro de la evaluación de las opiniones de Twitter en el método que propone, evaluando los intensificadores y modificadores de forma empírica.

CAPÍTULO 3. ESTADO DEL ARTE

En este capítulo se explicarán los trabajos relacionados con la comparación de los lexicones con respecto al análisis de sentimientos, tomando en cuenta el desarrollo del método y lexicón *VADER*. Con el objetivo de obtener que lexicones obtuvieron un mejor resultado de análisis de sentimientos.

Posteriormente se tomarán en cuenta la forma de evaluación de cada uno de los lexicones con respecto a la clasificación de emociones y polaridades de los documentos, textos u opiniones de redes sociales que los autores utilizaron para su análisis. Con el fin de obtener los mejores 7 lexicones que será utilizados para enriquecer el método basado en reglas *VADER*, explicando de forma detallada cada uno de los lexicones.

3.1. Trabajos relacionados

Los trabajos relacionados con respecto al estado del arte son los siguientes.

3.1.1. *VADER: A Parsomionous Rule-based Model for Sentiment Analysis of Social Media Text* (Hutto, 2014)

En el artículo presentado por Hutto y Gilbert en el año 2014 se desarrolla el método basado en reglas Diccionario de Valencias Conscientes para el Razonamiento de Sentimientos (*Value Aware Dictionary for sEntiment Reasoning VADER*, por sus siglas en inglés) con el objetivo de obtener una clasificación de polaridades y valencia de distintos documentos, textos y opiniones de redes sociales, editoriales del periódico de Nueva York, sinopsis de películas y opiniones de productos.

La idea que tuvo Hutto (2014) en su método es realizar un lexicón capaz de analizar características léxicas tales como abreviaturas, emoticones, acrónimos e iniciales. Hutto (2014) realiza una comparación entre los lexicones existentes clasificándolos en dos aproximaciones:

3.1.1.1. Lexicones de orientación semántica basada en polaridades

-LIWC

Lexicón validado por psicólogos, sociólogos y lingüistas compuesto de 4500 palabras divididas en 76 categorías, y 905 palabras divididas en categorías positivas y negativas. El lexicón de LIWC ha sido utilizado para analizar opiniones de Twitter en cuestiones políticas Tausczik (2010).

-GI

El lexicón de GI, contiene un total de 11,000 palabras clasificadas en 183 categorías. Para la construcción del lexicón de *VADER*, Hutto utilizó 1915 palabras positivas y 2291 palabras negativas del lexicón de GI (Stone, 1963).

-Hu-Liu (*opinion lexicon*)

Es un lexicón no supervisado basado en lexicones manuales y de diccionarios clasificando cada palabra de su diccionario mediante estudios lingüísticos. Las palabras del lexicón de Hu-Liu se obtuvieron de un conjunto de opiniones de productos. Evaluado mediante polaridades positivas y negativas, con un total de 6800 palabras, 2006 positivas y 4783 negativas (Hu, 2004).

3.1.1.2. Lexicones de intensidad basados en valencias

-ANEW

El lexicón de ANEW es desarrollado mediante el método no supervisado, conteniendo palabras obscenas, positivas, abreviaciones y acrónimos, agregadas de diversas páginas como *Original Balanced Affective Word List*¹⁸, *Urban Dictionary*¹⁹, *The Compass DeRose Guide to Emotion Words*²⁰, *Clustering words based on context similarity*²¹.

Contiene 1034 palabras evaluadas manualmente de acuerdo con la intensidad (valencia), con un intervalo de 1 a 9 con un punto neutral de 5, considerando como negativo la valencia menor a 5 y positivo las palabras con valencias mayores a 5 (Nielsen, 2016).

-SentiWordNet

Lexicón basado en diccionarios, se elaboró bajo el esquema supervisado y no supervisado, evaluando las polaridad positiva, negativa y neutral mediante un conjunto de datos de entrenamiento, dicho conjunto de entrenamiento está basado en sinónimos (explicado en el apartado 2 de esta tesis) determinando la valencia de cada palabra del lexicón.

Contiene un conjunto de 147,306 palabras etiquetadas con una valencia de 0 a 1 (Esuli, 2010).

-SenticNet

Desarrollado bajo el esquema supervisado, tomando en cuenta el lexicón de *SentiWordNet* para su elaboración. Contiene un total de 14,244 palabras etiquetadas con una valencia de -1 a 1.

¹⁸ <http://www.sci.sdsu.edu/CAL/wordlist/origwordlist.html>

¹⁹ <http://www.urbandictionary.com/>

²⁰ <http://www.derose.net/steve/resources/emotionwords/ewords.html>

²¹ <http://web-ngram.research.microsoft.com/similarity/>

Basado en la aplicación del procesamiento del lenguaje natural e inteligencia artificial el lexicón de *SenticNet* se estructura mediante conceptos (“hermoso día”, “risa”, “fiesta de cumpleaños”) y 16 emociones básicas llamadas el reloj de arena de las emociones (Cambria, 2010).

Hutto (2014) menciona que los lexicones anteriores tienden a ignorar características léxicas, es por ello que crea el lexicón de *VADER* a partir de tres reglas:

1. Extracción de la polaridad y valencia de palabras utilizando método supervisado y evaluación de humanos.
2. Elaboración y aplicación de reglas generadas con respecto a usos gramaticales y sintácticas.
3. Comparación del lexicón de *VADER* con respecto a otros lexicones.

Para la construcción del lexicón de *VADER*, Hutto (2014) realiza un lexicón base construido a partir de tres lexicones mencionados anteriormente (LIWC, ANEW y GI) agregando emoticones, acrónimos, iniciales y abreviaciones. Obteniendo así un lexicón base de 9,000 palabras.

Las palabras del lexicón base fueron evaluadas por un conjunto de 10 personas por palabra, mediante la interfaz gráfica *Amazon Mechanical Turk*²² (AMT), las palabras fueron evaluadas con una valencia en intervalo de -4 a 4 donde -4 es una valencia extremadamente negativa y +4 extremadamente positiva.

9 of 25

ROFL	Description: Rolling On Floor Laughing
------	---

[-1] Slightly Negative [-2] Moderately Negative [-3] Very Negative [-4] Extremely Negative

[0] Neutral (or Neither, N/A)

[1] Slightly Positive [2] Moderately Positive [3] Very Positive [4] Extremely Positive

Figura 1. Interfaz AMT.

²² <https://www.mturk.com/>

Con el objetivo de evitar que las palabras obtuvieran una evaluación incorrecta por parte de los evaluadores del AMT, Hutto (2014) aplicó reglas de control de calidad para la selección de las palabras del lexicon de *VADER*. Dichas reglas consisten en realizar una comparación entre los resultados otorgados por los evaluadores con respecto a un lexicon con resultados evaluados de manera correcta llamándolo *gold standard*.

Finalizado el lexicon con las valencias correctamente establecidas, Hutto (2014) obtuvo reglas gramaticales y sintácticas de un conjunto de 800 tweets evaluados mediante la herramienta *Pattern*²³, siendo las siguientes: signos de puntuación, capitalización de las palabras, modificadores, conjunciones y evaluación de negaciones. Incluyendo dichas reglas en el método de *VADER* con el fin de realizar análisis de sentimientos más preciso.

Una vez finalizado el método de *VADER* y su respectivo lexicon Hutto (2014) utiliza un corpus de opiniones de Twitter, opiniones de películas, opiniones de productos y opiniones de artículos del *New York Times*.

En la tabla 7 se muestran los resultados obtenidos por (Hutto, 2014).

²³ <http://pattern.com/>

Tabla 7. Resultados de VADER con respecto a otros lexicones (Hutto, 2014).

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)						Movie Reviews (10,605 review snippets)				
Ind. Humans	0.888	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	0.99	0.94	0.96	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)						NY Times Editorials (5,190 article snippets)				
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Se puede observar que el método *VADER* obtuvo un resultado de 0.96 para los Tweets. Concluyendo que el método y lexicón de *VADER* tiende a evaluar de forma eficaz las opiniones de Twitter.

3.1.2. Comparing and Combining Sentiment Analysis Methods (Goncalves, 2013)

En el trabajo siguiente, Goncalves (2013) menciona la importancia de las Redes Sociales (*Online Social Networks ONS* por sus siglas en inglés) y su análisis con respecto a diversos métodos y lexicones existentes ya sean basados en métodos supervisados y no supervisados. Clasificando las emociones con respecto a las polaridades.

Goncalves (2013) analiza dos conjuntos de datos, el primer conjunto es de 1.8 billones de tweets, y el segundo es un conjunto de textos evaluados por humanos.

El autor utilizó los lexicones de *LIWC*, *SentiWordNet* y *SenticNet*, explicados anteriormente. A continuación, se explican los lexicones diferentes utilizados por el autor.

-Emoticons

Este lexicón como su nombre lo indica es un diccionario evaluado mediante la polaridad positiva, negativa y neutral de emoticones, conteniendo alrededor de 80 emoticones (Read, 2005).

-SentiStrenght

Este lexicón está construido bajo el método supervisado tomando palabras de los lexicones LIWC y GI para su elaboración, teniendo 2310 palabras evaluadas mediante un intervalo de valencias de – 5 a 5 tomando como palabras positivas de 1 a 5 y palabras negativas de -1 a -5.

El lexicón de SentiStrength cuenta con características léxicas como: Eliminación de palabras repetidas, intensificadores, emoticones, signos de puntuación, eliminación de negaciones y capitalización (Thelwall, 2013).

-SASA

SailAil Sentiment Analyzer (SASA por sus siglas en inglés), es un método supervisado, evaluando las palabras de su lexicón mediante la interfaz AMT en polaridades positivas, negativas y neutrales (Wang, 2012).

-Happiness Index

El lexicón de *Happiness Index* se construyó a partir del lexicón de ANEW. Las palabras de su lexicón se encuentran evaluadas mediante un intervalo de 1 a 9 indicando la felicidad que existe en el texto. Siendo que las polaridades negativas se encuentran de 1 a 5 y las palabras positivas se encuentran entre 5 y 9 (Dodds, 2009).

-PANAS-t

El método PANAS-t así como su lexicón fueron creados por Goncalves (2013), con el fin de detectar las fluctuaciones de las emociones de los usuarios de Twitter. Las

palabras evaluadas en el lexicón de PANAS-t se encuentran en un intervalo de -1.0 a 1.0, obteniendo la polaridad positiva, negativa y neutral.

Con el objetivo de clasificar las emociones de los usuarios de Twitter con respecto a las polaridades del método PANAS-t, Goncalves (2013) realiza la siguiente aproximación.

En la tabla 8 se describe la clasificación de emociones con polaridades realizado por Goncalves (2013).

Tabla 8. Clasificación de emociones como polaridades (Goncalves, 2013).

Polaridad	Emoción
Positiva	Jovial, seguro, serenidad y sorpresa
Negativa	Miedo, tristeza, culpa, hostilidad, vergüenza y fatiga
Neutral	Atento

El primer conjunto de datos que utiliza Goncalves (2013) para su análisis se divide en seis eventos importantes dentro de Twitter, con el objetivo de obtener las opiniones más importantes.

En la tabla 9 se describen los eventos importantes que Goncalves (2013) toma en cuenta para su análisis.

Tabla 9. Selección de eventos y palabras clave de Twitter (Goncalves, 2013).

Eventos
AirFrance
2008US-Elect
2008Olimpics
Susan Boyle
H1N1
Harry Potter

El segundo conjunto de datos consistió en un set de seis conjuntos de datos evaluados por humanos mediante su polaridad positiva y negativa.

Se puede observar en la tabla 10 las opiniones extraídas por Goncalves (2013) de diferentes redes sociales, el total de opiniones y el porcentaje de palabras positivas y negativas del conjunto de opiniones.

Tabla 10. Conjunto de datos evaluados por humanos (Goncalves, 2013).

Tipo de datos	# Mensajes	Pos/Neg
Twitter	4,242	58.58%/41.42%
MySpace	1,041	84.17%/15.83%
YouTube	3,407	68.44%/31.56%
BBC fórum	1,000	13.16%/86.84%
Runners World	1,046	68.65%/31.35%
Digg	1,077	26.85%/73.15%

En la figura 2 se observan los resultados obtenidos por Goncalves (2013) con respecto al porcentaje de clasificación, donde se puede apreciar que *SentiWordNet* y *SenticNet* obtuvieron los mejores resultados.

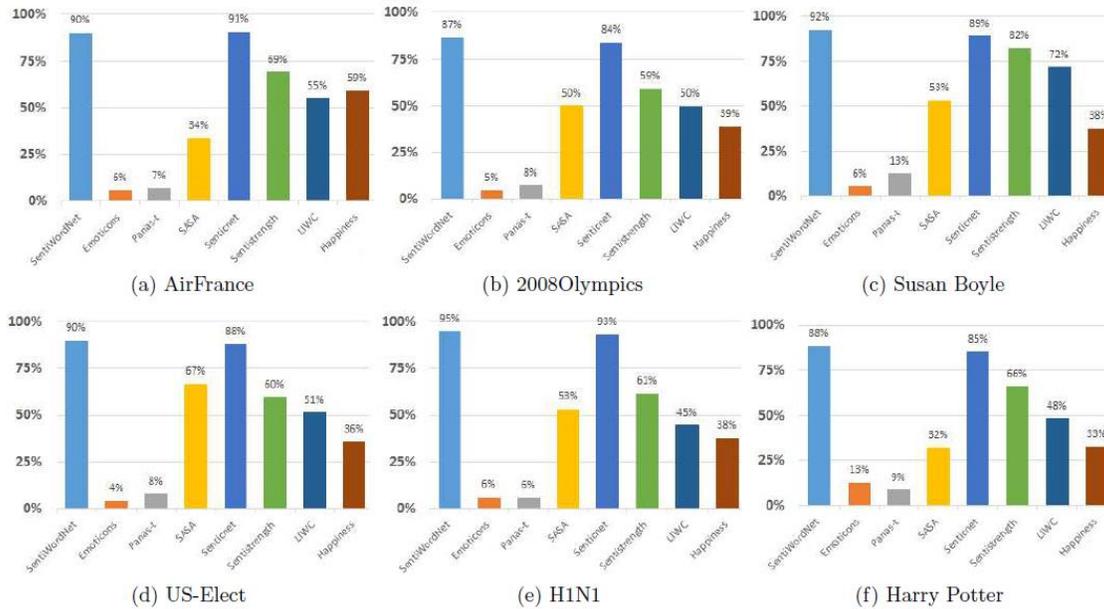


Figura 2. Comparación de resultados para opiniones de Twitter (Goncalves, 2013).

Para el segundo conjunto de datos se realizó la evaluación con respecto a las clasificaciones correctas de las polaridades positivas y negativas. Comparando que clasificación de la polaridad de los lexicones fuera igual a la clasificación de la polaridad del set de datos evaluado por los humanos, tomando en cuenta solamente los resultados obtenidos para las opiniones de twitter.

En la tabla 11 se muestran los dos mejores lexicones con sus respectivos resultados.

Tabla 11. Resultados de los dos mejores lexicones con respecto a la evaluación de opiniones de Twitter (Goncalves, 2013).

Método	Twitter
Emoticons	0.929
SentiStrength	0.843

Con el objetivo de obtener un mejor análisis de sentimientos Goncalves (2013) combina todos los lexicones, es decir las palabras y valencias de cada lexicon en uno solo.

En la figura 3 se muestra la evaluación para el resultado de los lexicones combinados.

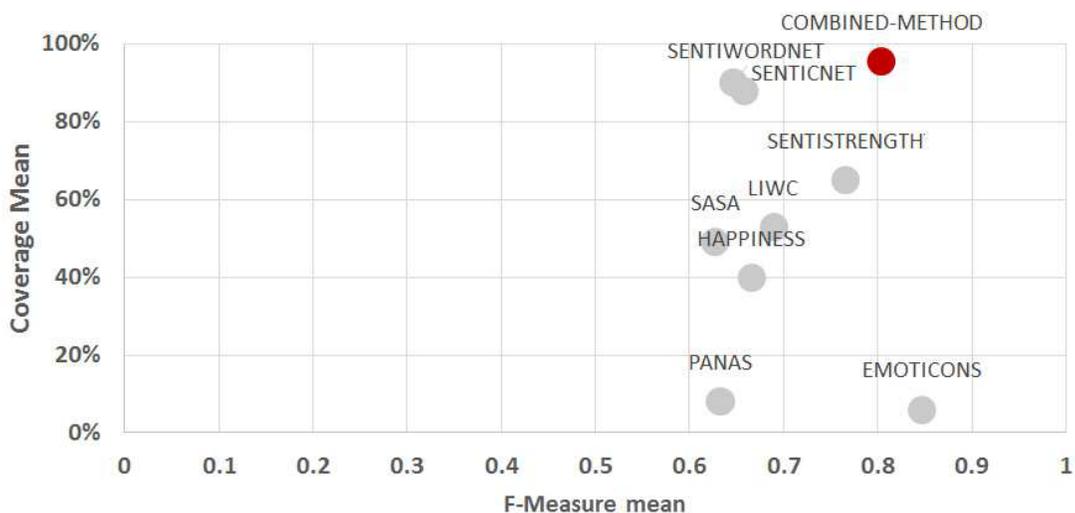


Figura 3. Resultados de la combinación de todos los lexicones (Goncalves, 2013).

En conclusión, se obtuvo que los mejores lexicones para el trabajo de Goncalves (2013) fueron *Emoticons*, *SentiStrength*, *SentiWordNet* y *SenticNet* en opiniones de

Twitter. Así como la combinación de todos los lexicones evaluados por el autor, obteniendo una evaluación superior a la de cada lexicón con un resultado de 0.81.

3.1.3. A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts (Musto, 2014)

En la investigación realizada por Musto (2014), propone encontrar el lexicón que obtenga la mejor evaluación al realizar análisis de sentimientos.

Con el objetivo de realizar la comparación de los lexicones Musto (2014) toma en cuenta el lexicón de *SentiWordNet*, *WordNet-Affect*, MPQA, *SenticNet*. Debido a que los lexicones MPQA y *WordNet-Affect* no han sido explicados en las investigaciones anteriores, se describirán con el objetivo de conocer cómo se estructuran.

-MPQA

Es un lexicón conformado por 8,222 palabras entre las cuales se encuentran los *Part Of Speech* (POS por sus siglas en inglés), los POS es la división de un documento, texto u opiniones en la estructura gramatical que lo conforman (sustantivos, pronombres, verbos, adverbios, adjetivos, conjunciones, preposiciones e intersecciones) ²⁴. Dichos POS se encuentran etiquetados mediante la polaridad positiva, negativa o neutral en el lexicón de MPQA (Wiebe, 2005).

-WordNet-Affect

La construcción del lexicón de *WordNet-Affect* hace una aproximación a los lexicones basados en diccionarios descritos en el capítulo 2 en el apartado de tipos de lexicón de esta tesis. Las palabras en lexicón de *WordNet-Affect* tienen por

²⁴ <http://partofspeech.org/>

nombre *A-LABEL*, estructurado de forma jerárquica evaluadas mediante su polaridad positiva, negativa y neutral (Strapparava, 2004).

La metodología que Musto (2014) desarrolló para obtener el resultado con respecto al análisis de sentimientos fue el dividir en micro-frases los tweets. Definiendo el sentimiento transmitido por un tweet como la suma de la polaridad transmitida por cada micro-frase. Musto (2014) etiquetó el resultado de la evaluación del sentimiento como $score(tj)$. Tomando en cuenta que las negaciones en las opiniones de Twitter tienden a invertir el resultado del sentimiento de las opiniones.

Con el fin de calcular el sentimiento y la polaridad de las opiniones, el autor elabora cuatro aproximaciones: básico, normalizado, enfatizado y normalizado-enfatizado.

Dichas aproximaciones dependen del cálculo del sentimiento $score(tj)$ es por ello qué para cada lexicón se obtuvo realizando la comparación de las palabras de los lexicones con respecto a las opiniones de los corpus utilizados por Musto (2014), si las palabras de los lexicones son iguales a las palabras escritas en las opiniones, se realiza una suma de las valencias para cada opinión obteniendo así $score(tj)$.

El primer corpus que utilizó Musto (2014) fue de SemEval del año 2013, el cual está compuesto por 14,435 Tweets, siendo manualmente evaluados con respecto a su polaridad positiva, negativa y neutral. El segundo corpus *Stanford Twitter Sentiment* (STS) contiene más de 1,600,000 Tweets evaluando su polaridad mediante los emoticones que contiene cada opinión.

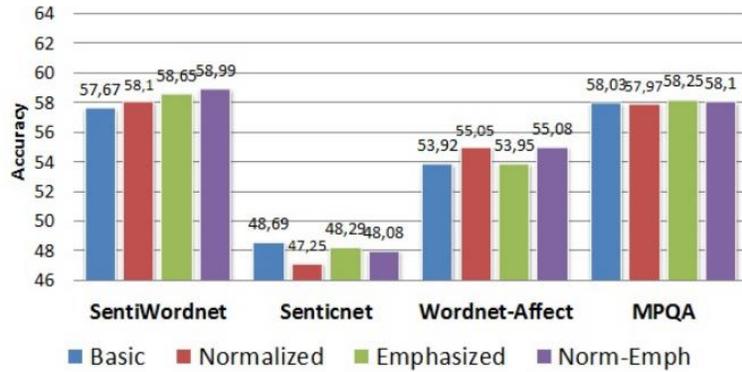


Figura 4. Evaluación para el corpus de SemEval (Musto, 2014).

Se puede observar que para el corpus de SemEval del año 2013 Musto (2014) obtiene el mejor resultado para el lexicón de *SentiWordNet*.

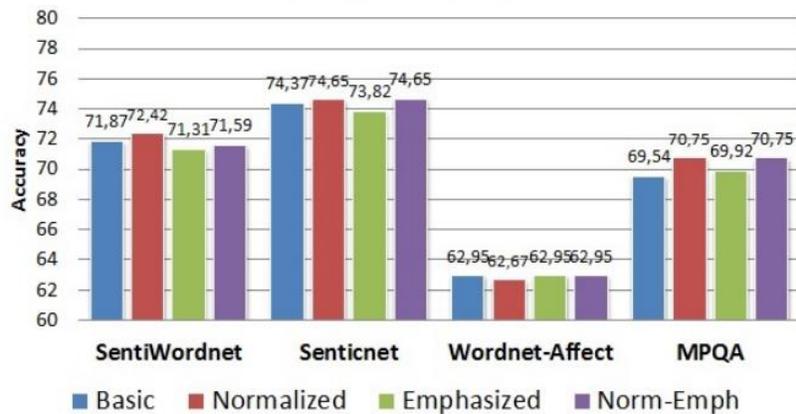


Figura 5. Evaluación para el corpus STS 2013 (Musto, 2014).

Para el corpus de STS Musto (2014) probó que el lexicón de *SenticNet* fue el mejor en comparación a los lexicones utilizados.

Concluyendo la explicación de la investigación de Musto (2014) se obtiene que los mejores lexicones son *SentiWordNet* y *SenticNet*, que, al igual que en la investigación de Goncalves (2013), obtuvo dichos lexicones como los mejores.

3.1.4. Enhancing Lexicon-Based Review Classification by Merging and Revising Sentiment Dictionaries (Cho, 2013)

Debido a la creciente investigación con respecto a análisis de sentimientos, Cho (2013) menciona que se han construido diversos recursos léxicos con el objetivo de cubrir dicho crecimiento.

En los lexicones evaluados por Cho (2013) se encuentran 3 diferentes lexicones con respecto a los mencionados en las investigaciones anteriores (*General Inquirer*, *Hu-Liu opinion lexicon*, *SentiWordNet*, *WordNet-Affect*, *MPQA*), describiendo los 4 nuevos lexicones que utilizó el autor para su evaluación.

-AFINN

Contiene 2,477 palabras en su lexicón evaluados mediante un rango de valencia de -5 a 5 donde -5 se considera como muy negativo y 5 como muy positivo²⁵.

-Micro-WNOp

Constituido por 2,190 sinónimos y 1,960 palabras, evaluadas mediante polaridades positivas y negativas. La elaboración del lexicón Micro-WNOp hace una aproximación a los lexicones basados en diccionarios²⁶.

-SentiSense

Es un lexicón basado en diccionarios, evaluando emociones tales como “felicidad”, “amor”, “esperanza”, “calma”, y “gusto” con una valencia de 0 a 1.0. “Miedo”, “furia”,

²⁵ http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

²⁶ <http://www-3.unipv.it/wnop/>

“disgusto”, “sorpresa” y “anticipación” se evaluaron con una valencia de 0 a – 1.0 y “ambiguo”, “sorpresa” y “anticipación” con una valencia de 0²⁷.

Con el fin de evaluar los lexicones anteriormente mencionados, Cho (2013) toma en cuenta un conjunto de datos de 90,000 opiniones de libros de Amazon. Para obtener la valencia de cada opinión ($RSS(D_j)$) del conjunto de datos de Amazon, Cho (2013) aplicó la siguiente fórmula.

$$RSS(D_j) = \frac{1}{n} \sum_{i=1}^n D_j(w_i) \quad (6)$$

Donde $D_j(w_i)$ son las palabras evaluadas de los lexicones utilizados para el análisis y n es el total de palabras encontradas en las opiniones que coinciden con las palabras de cada lexicon.

Una vez obtenida la valencia de cada opinión, se le asignó una polaridad positiva o negativa a cada opinión de Amazon.

Debido a que los lexicones no contienen las mismas evaluaciones (valencias, polaridades o emociones) en las palabras que contienen, Cho (2013) realiza una estandarización donde:

- Las polaridades positivas, negativas o neutrales fueron iguales a las a valencias de – 1.0 como polaridad negativa, 0.0 como neutral y 1.0 como positivo.
- Las emociones como felicidad, tristeza, amor, etc. Fueron mapeadas con respecto a la polaridad positiva, negativa o neutral que indican cada emoción.

Una vez obtenidas las polaridades estandarizadas las opiniones de Amazon y los lexicones, Cho (2013) obtiene los siguientes resultados. En la tabla 12 se muestran los resultados obtenidos por Cho (2013), así como los lexicones utilizados.

²⁷ <http://nlp.uned.es/~jcalbornoz/SentiSense.html>

Tabla 12. Evaluación de la exhaustividad y precisión de los lexicones Cho (2013).

Diccionario	Resultado
AFINN	63.8%
General Inquirer	65.8%
Micro-WNOp	55.4%
Opinion Lexicon	65.4%
SentiSense	62.8%
<i>SentiWordNet</i>	65.7%
Subjectivity Lexicon	64.7%
<i>WordNet-Affect</i>	49.0%

Con el objetivo de mejorar la precisión de la clasificación de las opiniones de Amazon, Cho (2013) realiza una combinación de los lexicones excluyendo a los peores evaluados. Realizando una combinación en pares de los lexicones obteniendo un total de 21 combinaciones, una vez analizados los 21 pares de combinaciones Cho (2013) realiza una combinación del conjunto de todos los lexicones.

En la figura 6 se muestra el resultado de la precisión al combinar los lexicones, donde cada línea representa la fusión de los mejores lexicones.

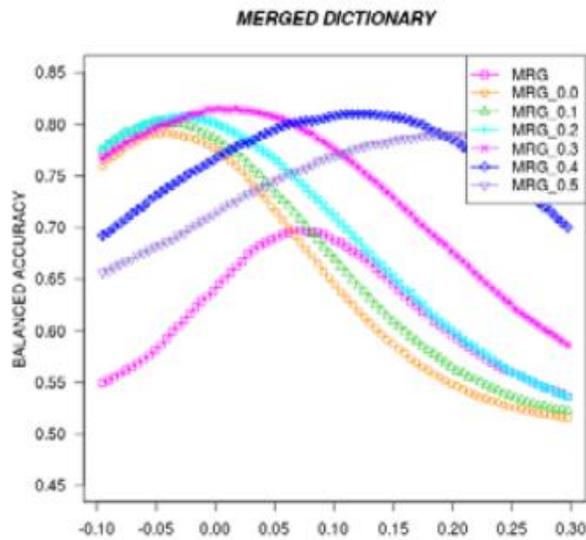


Figura 6. Fusión de los lexicones.

En la investigación de Cho (2013), se puede concluir que los mejores lexicones para opiniones de Amazon fueron *General Inquirer*, *SentiWordNet* y *Hu-Liu*.

3.1.5. *Lexicon-Based Methods for Sentiment Analysis* (Taboada, 2011)

Con el objetivo de comparar el método y lexicon SO-Cal (*Semantic Orientation Calculator*) Taboada (2011) carga diversos lexicones en el método SO-Cal.

Para ello Taboada (2011) utiliza los lexicones de *Google-generated PMI-based dictionary*, *Maryland*, *General Inquirer*, *Subjectivity Lexicon*, *SentiWordNet* y *SO-Cal*. En investigaciones anteriores se describieron los lexicones *General Inquirer*, *Subjectivity Lexicon* y *SentiWordNet*, es por ello que se explicarán los diferentes lexicones.

-Google-generated PMI-based dictionary

Lexicón creado mediante Google, evaluado mediante polaridades (positivas, negativas y neutrales) y valencias con un intervalo de -5 a 5²⁸.

-Maryland

El lexicón de Maryland está compuesto por alrededor de 70,000 palabras evaluadas mediante polaridades positivas y negativas, con un intervalo de valencias de -3 a 3²⁹.

-SO-Cal Lexicón

El lexicón de SO-Cal está desarrollado mediante el método no supervisado. Contiene más de 6000 palabras clasificadas mediante un rango de valencia de -5 a 5. Creado a creado a partir de pequeños lexicones entre los cuales se encuentran:

- Adjetivos: Obtenidos mediante los lexicones de GI y *WordNet*.
- Sustantivos, verbos y adverbios. Es un pequeño lexicón con una valencia en sus palabras de -5 a 5 estructurado por 2,252 adjetivos, 1,142 sustantivos, 903 verbos y 745 adverbios.
- Intensificadores: Utilizados para incrementar la orientación semántica del análisis de sentimientos, este pequeño lexicón contiene 177 intensificadores entre ellos signos de exclamación, mayúsculas y conectores.
- Negaciones: Conjunto de negaciones que tiene el objetivo de invertir el valor de la orientación semántica (Taboada, 2011).

Con el fin de obtener las evaluaciones de los lexicones en el método de SO-Cal, Taboada (2011) analizó diversos corpus entre ellos opiniones de películas, cámaras y opiniones de redes sociales.

²⁸ <http://www.pmgloss.com/about/>

²⁹ <https://clipdemos.umiacs.umd.edu/englcslex/>

En la tabla 13 se muestran los resultados con respecto a la evaluación de los corpus utilizados por Taboada (2011) mediante el método SO-Cal y lexicones utilizados.

Tabla 13. Evaluación porcentual de cada lexicon cargado en el método SO-Cal (Taboada, 2011).

Diccionario	Porcentaje correcto por corpus				
	Opiniones 1	Opiniones 2	Películas	Cámaras	Global
Google Lexicón	62.00	58.50	66.31	61.25	62.98
Maryland	58.00	63.75	67.42	59.46	62.65
GI	68.00	70.50	64.21	72.33	68.02
SentiWordNet	66.50	66.50	61.89	67.00	65.02
Subjetivity	72.75	71.75	65.42	77.21	72.04
SO-Cal	80.25	80.00	76.37	80.16	78.74

El lexicon que obtuvo un mayor porcentaje de evaluación fue el realizado por Taboada (2011), SO-Cal.

3.2. Resumen del capítulo

En el estado del arte se logró obtener de las diferentes investigaciones los lexicones utilizados por los autores y la metodología que utilizaron para obtener el mejor lexicon evaluado.

Tabla 14. Mejores lexicones obtenidos.

Autor	Mejores lexicones	Evaluación
(Hutto, 2014)	VADER	0.96
(Goncalves, 2013)	Emoticons	0.92
	SentiStrength	0.84
(Musto, 2014)	SentiWordNet	58.99
	SenticNet	74.65
(Cho, 2013)	General Inquirer	65.8
	Hu-Liu (opinion lexicon)	65.2
(Taboada, 2011)	SO-Cal	78.74

Se puede observar en la tabla 14 los mejores lexicones obtenidos de los trabajos explicados en los párrafos anteriores.

Debido a que los lexicones *General Inquirer* y *Emoticons* evaluados en los trabajos de Goncalves (2013) y Cho (2013) no se encuentran etiquetados con respecto a la valencia, se utilizarán los segundos lexicones evaluados en dichos trabajos.

Tomando en cuenta los siguientes 7 lexicones para enriquecer el método *VADER*:

- VADER
- SentiStrength
- SentiWordNet
- SenticNet
- Hu-Liu (*opinion lexicon*)
- SO-Cal
- ANEW

El lexicon de ANEW fue elegido ya que se desea comprobar su efectividad en el método de *VADER*, debido a las características anteriormente descritas del lexicon.

Una vez obtenidos los mejores lexicones de cada uno de los trabajos de los autores, se realizará una aproximación con respecto al trabajo de Taboada (2011), Goncalves (2013) y Cho (2013). Taboada (2011) propone evaluar mediante el método SO-Cal cada uno de los lexicones que utilizó, realizando dicha aproximación con respecto al método *VADER* y los mejores lexicones obtenidos.

Con respecto al trabajo de Goncalves (2013) y Cho (2013) se combinarán los lexicones enlistados anteriormente, realizando una combinación en pares y una combinación total de los lexicones con el objetivo de enriquecer el método basado en reglas *VADER*.

CAPÍTULO 4. MARCO METODOLÓGICO

En este capítulo se describen las etapas de la metodología propuesta con el objetivo de enriquecer el método basado en reglas *VADER*.

4.1. Metodología propuesta

En el capítulo 1 de esta tesis se formuló la problemática a resolver ¿Cómo enriquecer el método basado en reglas (*VADER*) a través de lexicones para un corpus de Twitter en idioma inglés? Teniendo en cuenta que el método *VADER* y los lexicones a utilizar para enriquecer el método *VADER* analizan únicamente tweets en idioma inglés, es necesario obtener un corpus de tweets en dicho idioma.

Para poder resolver dicho problema se propone la siguiente metodología, en la figura 7 se puede observar los pasos para resolver la problemática mencionada en el párrafo anterior.

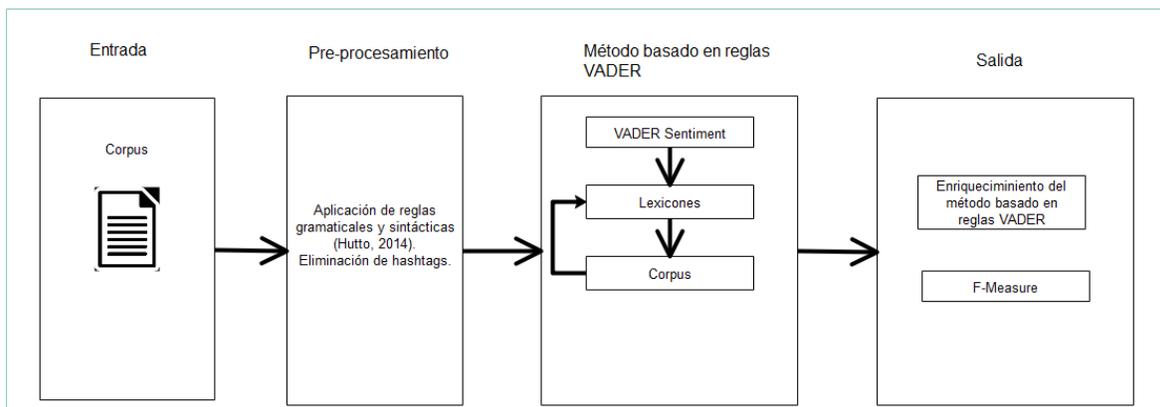


Figura 7. Etapas del método propuesto.

A continuación, se describen las etapas de la metodología propuesta.

4.1.1. Entrada

En la primera etapa de la metodología propuesta se tiene como objetivo obtener corpus de opiniones de Twitter en idioma inglés, recordando que en el capítulo 1 de la investigación en el apartado de introducción se mencionaron algunos de los concursos que se llevan a cabo a nivel mundial con el fin de obtener el mejor método para análisis de sentimientos; se tomó en cuenta el corpus del concurso de WASSA del año 2017, obteniendo un conjunto de datos de entrenamiento y de pruebas.

El corpus de opiniones de Twitter en idioma inglés del concurso de WASSA del año 2017 se divide en 4 emociones felicidad, tristeza, ira y miedo (Balahur, 2017). Con el objetivo de clasificar las emociones con respecto a las polaridades del método *VADER*, se realiza una clasificación como la de Goncalves (2013) y Cho (2013) con respecto a las emociones como polaridades.

Obteniendo la siguiente clasificación:

Felicidad=Positivo.

Furia, Tristeza, Miedo= Negativo, Neutral.

En el anexo B.1 se muestra un ejemplo del corpus utilizado.

4.1.2. Preprocesamiento

En este paso de la metodología se eliminarán los hashtags de las opiniones de Twitter con el objetivo de obtener una clasificación de opiniones más precisa, en los hashtags de Twitter comúnmente se encuentran palabras que expresan una emoción o sentimiento con respecto a la opinión escrita.

Con respecto a lo explicado en el trabajo de Hutto (2014), el método evalúa las características gramaticales y sintácticas de las opiniones de Twitter, es por ello que

se realizará el preprocesamiento en las opiniones de Twitter con respecto a dichas reglas. En el anexo B.2 se muestra un ejemplo de las opiniones preprocesadas.

4.1.3. Método basado en reglas *VADER*

El método para realizar la clasificación de emociones a través de las polaridades de *VADER* se realiza a través de diversas etapas.

4.1.3.1. *VADER Sentiment*

El método basado en reglas *VADER* explicado en el capítulo 3 en el trabajo de Hutto (2014) de esta tesis se aplicará con el objetivo de obtener la clasificación de emociones del corpus de Twitter de WASSA del año 2017.

El método *VADER* se desarrolló en el lenguaje de programación Python, utilizando diversos comandos para su instalación, así como las librerías `grpcio`³⁰ y `requests`³¹ (librería utilizadas para la conexión entre cliente y servidor). En el anexo A.1 se muestran los comandos utilizados para su instalación.

Con el objetivo de obtener una clasificación por opinión de Twitter, Hutto (2014) realiza una normalización parecida a Musto (2014) y Cho (2013) de las palabras encontradas en la opinión de Twitter con respecto a las evaluadas en el lexicon utilizado. Dicha normalización la realiza de la siguiente manera.

$$\frac{x}{\sqrt{x^2+a}} \quad (7)$$

³⁰ <https://grpc.io/docs/tutorials/basic/python.html>

³¹ <http://docs.python-requests.org/en/master/>

Donde x es la suma de las valencias de las palabras de los lexicones encontradas en los tweets y a es un parámetro de normalización que Hutto (2014) establece.

Obteniendo por opinión normalizada la polaridad con un intervalo de -1 a 1, donde -1 es negativo, 0 es neutral y 1 positivo.

4.1.3.2. Lexicones

Como se había mencionado, en el trabajo de Goncalves (2013) y Cho (2013) realizan una combinación de los lexicones formando grupos de dos lexicones para la evaluación. Para el caso de la tesis presentada, de los 7 lexicones a evaluar se tiene una combinación de 21 pares de lexicones que comprobar con el objetivo de encontrar los lexicones que enriquecerán al método *VADER*.

La tabla 15 muestra la combinación a realizar por cada lexicón y el número de palabras que conforman a cada combinación.

Tabla 15. Total de palabras con respecto a la combinación de lexicones en pares.

Lexicones	VADER	SentiStrength	SentiWordNet	SenticNet	Hu-Liu	SO-Cal	ANEW
VADER		10,179	58,059	107,517	21,091	13,763	8,547
SentiStrength	10,179		53,204	102,662	16,236	8,908	3692
SentiWordNet	58,059	53,204		150,542	64,116	56,788	51,572
SenticNet	107,517	102,662	150,542		113,574	106,246	101,030
Hu-Liu	21,091	16,236	64,116	113,574		19,820	14,604
SO-Cal	13,763	8,908	56,788	106,246	19,820		7,276
ANEW	8,547	3,692	51,572	101,030	14,604	7,276	

De igual manera se requiere saber si la combinación de todos los lexicones resulta en un enriquecimiento mayor en el método *VADER* obteniendo un lexicon con un total de 181,569 palabras. Así como la evaluación individual de los lexicones. En el anexo A.1 al A.8 se muestra un ejemplo de los lexicones evaluados.

4.1.3.3. Corpus

Se carga el corpus de WASSA del año 2017 de Twitter en el método *VADER*, obteniendo así el valor de la normalización por cada opinión.

4.1.4. Salida

Se muestra los resultados obtenidos para cada uno de los pares de lexicones combinados.

4.1.4.1. Enriquecimiento del método basado en reglas *VADER*

Con el objetivo de enriquecer el método de *VADER* se realizarán la etapa de carga de las 21 combinaciones de los lexicones, encontrando que par de lexicones enriquecen el método *VADER*, al final se cargará la combinación de todos los lexicones y comprobar los resultados con respecto a las combinaciones en pares.

4.1.4.2. *F-Measure*

De acuerdo con el trabajo de Goncalves (2013), evalúa el desempeño de los lexicones con respecto a la precisión (*precision*), exhaustividad (*recall*) y medida F

(*F-Measure*), realizando las ecuaciones correspondientes. Donde *a* son las opiniones positivas correctas, *b* las opiniones negativas clasificadas como positivas, *c* las opiniones positivas clasificadas como negativas y *d* las opiniones negativas clasificadas correctas. En la tabla 16 se muestra de manera simplificada lo anterior explicado.

Tabla 16. Clasificación de opiniones con respecto a su polaridad.

		Positivos	Negativos
Expectativa de predicción	Positivos	<i>a</i>	<i>b</i>
	Negativos	<i>c</i>	<i>d</i>

$$R = a/(a + c) \quad (8)$$

$$P = a/(a + b) \quad (9)$$

$$F \text{ Measure} = 2 * (P * R)/(P + R) \quad (10)$$

De acuerdo con la clasificación de emociones con respecto a polaridades, se calculará la medida F con respecto a las opiniones clasificadas *a*, *b*, *c* y *d* por el método *VADER* con respecto a las ya clasificadas en el corpus de WASSA del año 2017. En el anexo C.3 se muestra el script desarrollado para calcular la medida F.

CAPÍTULO 5. EXPERIMENTOS Y RESULTADOS

En este capítulo se muestran los experimentos realizados con respecto a la evaluación de los lexicones individuales, combinados en pares y todos combinados.

Con el objetivo de mostrar el enriquecimiento del método basado en reglas *VADER* y su eficacia con respecto a la clasificación de emociones, se realizará la experimentación en dos fases. La primera fase será comprobar el método *VADER* y la evaluación de los lexicones individuales, así como la combinación de los lexicones sin realizar el preprocesamiento mencionado en el capítulo anterior. La segunda fase será comprobar el método *VADER* y la evaluación de los lexicones individuales, así como las combinaciones de los lexicones realizando el preprocesamiento.

5.1. Experimentación sin preprocesamiento

5.1.1. Lexicones individuales

Se evalúan los lexicones de manera individual obteniendo el resultado de la clasificación de emociones mediante la precisión, exhaustividad y medida F, teniendo en consideración que el resultado más cercano de 1 es el mejor lexicon evaluado.

Tabla 17. Resultados obtenidos por cada lexicón cargado en el método basado en reglas VADER.

Lexicones	Medida F	Precisión	Exhaustividad
VADER	0.54	0.42	0.75
SentiStrength	0.35	0.35	0.35
SentiWordNet	0.36	0.23	0.8
SenticNet	0.41	0.26	0.87
Hu-Liu	0.51	0.4	0.7
SO-Cal	0.52	0.41	0.72
ANEW	0.35	0.25	0.63

En la tabla 17 se muestran los mejores lexicones evaluados con un valor mayor a 0.50 con respecto a la medida F, siendo los siguientes lexicones:

- VADER
- SO-Cal
- Hu-Liu

5.1.2. Lexicones combinados en pares

Con el objetivo de enriquecer el método basado en reglas se realiza la combinación de los lexicones, mostrando los resultados con respecto a la precisión, exhaustividad y medida F.

En la tabla 18 se muestran los mejores lexicones evaluados con respecto a la precisión con una evaluación mayor a 0.40.

Tabla 18. Resultados obtenidos con respecto a la precisión de la combinación de los lexicones.

Lexicones	VADER	SentiStrength	SentiWordNet	SenticNet	Hu-Liu	SO-Cal	ANEW
VADER		0.42	0.3	0.27	0.38	0.41	0.28
SentiStrength	0.42		0.25	0.26	0.38	0.41	0.26
SentiWordNet	0.3	0.25		0.26	0.32	0.33	0.23
SenticNet	0.27	0.26	0.26		0.29	0.3	0.24
Hu-Liu	0.38	0.38	0.32	0.29		0.41	0.27
SO-Cal	0.41	0.41	0.33	0.3	0.41		0.27
ANEW	0.28	0.26	0.23	0.24	0.27	0.27	

En la tabla 19 se muestran los mejores lexicones evaluados con respecto a la exhaustividad con una evaluación mayor o igual a 0.80.

Tabla 19. Resultados obtenidos con respecto a la exhaustividad de la combinación de los lexicones.

Lexicones	VADER	SentiStrength	SentiWordNet	SenticNet	Hu-Liu	SO-Cal	ANEW
VADER		0.75	0.84	0.88	0.76	0.8	0.87
SentiStrength	0.75		0.77	0.87	0.71	0.73	0.73
SentiWordNet	0.84	0.77		0.88	0.83	0.82	0.87
SenticNet	0.88	0.87	0.88		0.89	0.87	0.93
Hu-Liu	0.76	0.71	0.83	0.89		0.76	0.83
SO-Cal	0.8	0.73	0.82	0.87	0.76		0.84
ANEW	0.87	0.73	0.87	0.93	0.83	0.84	

En la tabla 20 se muestran los mejores lexicones evaluados con respecto a la medida F con una evaluación mayor o igual a 0.50.

Tabla 20. Resultados obtenidos con respecto a la medida F de la combinación de los lexicones.

Lexicones	VADER	SentiStrength	SentiWordNet	SenticNet	Hu-Liu	SO-Cal	ANEW
VADER		0.54	0.44	0.41	0.51	0.54	0.42
SentiStrength	0.54		0.38	0.40	0.50	0.52	0.38
SentiWordNet	0.44	0.38		0.40	0.46	0.48	0.36
SenticNet	0.41	0.40	0.40		0.43	0.45	0.39
Hu-Liu	0.51	0.51	0.46	0.43		0.53	0.41
SO-Cal	0.54	0.52	0.48	0.45	0.53		0.41
ANEW	0.42	0.38	0.36	0.39	0.41	0.41	

Así como en los resultados de los lexicones individuales, se toman en cuenta los lexicones con una evaluación de medida F superior a 0.50.

Donde los mejores pares de lexicones que enriquecen al método basado en reglas son:

- VADER/SentiStrength
- VADER/Hu-Liu
- VADER/SO-Cal
- SentiStrength/Hu-Liu
- SentiStrength/SO-Cal
- Hu-Liu/SO-Cal

5.1.3. Lexicones combinados

A continuación, se realizará la combinación de los 7 lexicones y su respectivo resultado, tomando en cuenta que, para obtener los resultados de los lexicones individuales y combinados se realizó la interfaz de la figura 8.

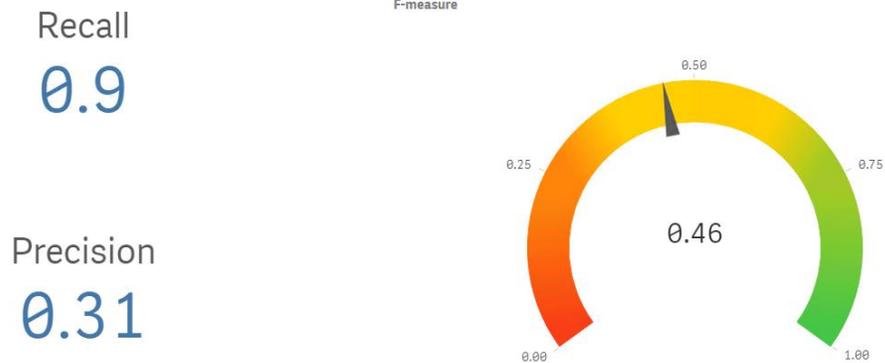


Figura 8. Interfaz de resultado de lexicones combinados.

Se puede apreciar que la combinación de todos los lexicones obtiene un resultado de medida $F = 0.46$ el cual no mejora la clasificación de emociones y el enriquecimiento del método basado en reglas *VADER* con respecto a los resultados anteriores.

5.2. Experimentación con preprocesamiento

5.2.1. Lexicones individuales

Los resultados de los lexicones individuales realizando el preprocesamiento del corpus con respecto a la medida F , precisión y exhaustividad son los siguientes.

Tabla 21. Resultados de los lexicones individuales realizando el preprocesamiento del corpus de opiniones.

Lexicones	Medida F	Precisión	Exhaustividad
VADER	0.74	0.66	0.84
SentiStrength	0.49	0.59	0.42
SentiWordNet	0.36	0.23	0.82
SenticNet	0.45	0.3	0.91
Hu-Liu	0.64	0.54	0.79
SO-Cal	0.66	0.57	0.79
ANEW	0.35	0.23	0.71

En la tabla 21 se muestra el mejor lexicón evaluado con un valor mayor a 0.70 con respecto a la medida F:

- *VADER*

5.2.2. Lexicones combinados en pares

A continuación, se muestran los resultados para los lexicones combinados en pares.

En la tabla 22 se muestran los mejores lexicones evaluados con respecto a la precisión con una evaluación mayor o igual a 0.60.

Tabla 22. Resultados obtenidos con respecto a la precisión.

Lexicones	VADER	SentiStrength	SentiWordNet	SenticNet	Hu-Liu	SO-Cal	ANEW
VADER		0.66	0.43	0.34	0.56	0.61	0.29
SentiStrength	0.66		0.3	0.32	0.55	0.6	0.25
SentiWordNet	0.43	0.3		0.29	0.42	0.45	0.22
SenticNet	0.34	0.32	0.29		0.35	0.4	0.24
Hu-Liu	0.56	0.55	0.42	0.35		0.57	0.27
SO-Cal	0.61	0.6	0.45	0.4	0.57		0.28
ANEW	0.29	0.25	0.22	0.24	0.27	0.28	

En la tabla 23 se muestran los mejores lexicones evaluados con respecto a la exhaustividad con una evaluación mayor o igual a 0.90.

Tabla 23. Resultados obtenidos con respecto a la exhaustividad.

Lexicones	VADER	SentiStrength	SentiWordNet	SenticNet	Hu-Liu	SO-Cal	ANEW
VADER		0.85	0.89	0.91	0.86	0.86	0.92
SentiStrength	0.85		0.8	0.9	0.8	0.79	0.79
SentiWordNet	0.89	0.8		0.91	0.88	0.88	0.9
SenticNet	0.91	0.9	0.91		0.93	0.91	0.95
Hu-Liu	0.86	0.8	0.88	0.93		0.83	0.89
SO-Cal	0.86	0.79	0.88	0.91	0.83		0.9
ANEW	0.92	0.79	0.9	0.95	0.89	0.9	

En la tabla 24 se muestran los mejores lexicones evaluados con respecto a la medida F con una evaluación mayor a 0.70.

Tabla 24. Resultados de la combinación los lexicones con respecto al resultado de la medida F.

Lexicones	VADER	SentiStrength	SentiWordNet	SenticNet	Hu-Liu	SO-Cal	ANEW
VADER		0.74	0.58	0.50	0.68	0.72	0.44
SentiStrength	0.74		0.44	0.47	0.65	0.68	0.38
SentiWordNet	0.58	0.44		0.44	0.57	0.60	0.36
SenticNet	0.50	0.47	0.44		0.51	0.56	0.39
Hu-Liu	0.68	0.65	0.57	0.51		0.67	0.42
SO-Cal	0.72	0.68	0.60	0.56	0.67		0.42
ANEW	0.44	0.38	0.36	0.39	0.42	0.42	

El resultado de la combinación de los lexicones con respecto al corpus de Twitter preprocesado tuvo un mayor enriquecimiento del método con respecto a la clasificación de emociones, tomando en cuenta la combinación de lexicones con un valor de medida F mayor a 0.70.

Donde los mejores pares de lexicones que enriquecen al método basado en reglas son:

- VADER/SentiStrength
- VADER/SO-Cal

5.2.3. Lexicones combinados

Se muestra la evaluación de los 7 lexicones combinados con respecto al corpus preprocesado.

Resultados

Recall
0.94

Precision
0.43

F-measure



Figura 9. Interfaz de resultado de lexicones combinados con preprocesamiento.

Se puede observar que la combinación de todos los lexicones del corpus preprocesado obtiene un resultado de medida $F = 0.59$ el cual mejora con respecto a la evaluación sin realizar el preprocesamiento.

CAPÍTULO 6. CONCLUSIONES Y TRABAJO FUTURO

En la presente tesis se cumplieron los siguientes objetivos.

Se analizó el método basado en reglas (*VADER*) con respecto al trabajo de Hutto (2014) presentando las reglas que aplicó para el desarrollo del método y el lexicon *VADER*, obteniendo buenos resultados al analizar tweets en su investigación.

Se evaluó el desempeño del método *VADER* cargando 7 diversos lexicones obtenidos de las investigaciones vistas en el capítulo 3 para la clasificación de emociones del corpus de WASSA del año 2017, donde el lexicon de *VADER* obtuvo un resultado de 0.74 y SO-Cal con 0.66, dichos lexicones enriquecieron al método *VADER*.

Con el objetivo de verificar si el método *VADER* era enriquecido con respecto a la combinación de lexicones que Goncalves (2013) y Cho (2013) realizaron en sus respectivos trabajos, se combinó los 7 lexicones. Los mejores resultados se obtuvieron en la combinación de lexicones en pares, donde las mejores combinaciones fueron *VADER/SentiStrength* con un resultado de 0.74 y *VADER/SO-Cal* con un resultado del 0.72. Se pudo observar que la combinación *VADER/SO-Cal* enriqueció mejor al método *VADER* que el lexicon individual *SO-Cal*. Con respecto a la combinación de los 7 lexicones no se observó una mejoría en el enriquecimiento del método *VADER* obteniendo un resultado de 0.59.

Con el fin de obtener los mejores resultados tanto de los lexicones evaluados de forma individual y combinados se realizó el preprocesamiento del corpus de WASSA de año 2017, eliminando los hashtags y aplicando las reglas gramaticales y sintácticas en los tweets que Hutto (2014) realizó en su trabajo. Evaluando los resultados para cada tweet del corpus de WASSA del año 2017 con respecto a los obtenidos por el método *VADER*. En el anexo B.1 y B.2 se puede observar el preprocesamiento realizado de corpus.

6.1. Trabajo Futuro

Obtener una mejor clasificación de emociones mediante el método basado en reglas y combinación de lexicones.

Desarrollar un lexicón que obtenga mejores resultados para el método basado en reglas *VADER*.

Elaborar un lexicón en idioma español en el método basado en reglas *VADER*, con el fin de realizar análisis de sentimientos en opiniones de Twitter en español.

REFERENCIAS

- (Aguado, 2010) Aguado de Cea Guadalupe, Barrios M. Auxiliadora, Bernardos M. Socorro, Campanella Inés, Montiel Ponsoda Elena, Muñoz García Óscar, Rodríguez Victor, Análisis de sentimientos en un corpus de redes sociales, 2010 http://oa.upm.es/20092/1/INVE_MEM_2012_137811.pdf.
- (Alberich, 2007) Alberich, Mario. (2007). Procesamiento del Lenguaje Natural Guía Introductoria. 2018, de ninguno Sitio web: <https://sopadebits.com/wp-content/uploads/2011/03/4479-pln-1.0-20070630.pdf>
- (Amores, 2016) Amores, Mario & Arco, Leticia & Barrera, Abel. (2016). Efectos de la Negación, Modificadores, Jergas, Abreviaturas y Emoticonos en el Análisis de Sentimiento. 2018, de CEUR Workshop Proceedings Sitio web: <http://ceur-ws.org/Vol-1797/paper5.pdf>
- (Apidianaki, 2018) Apidianaki, Marianna & Mohammad, Saif & May, Jonathan. (2018). International Workshop on Semantic Evaluation. 2018, de Codalab Sitio web: <http://alt.qcri.org/semeval2018/index.php?id=codalab>
- (Balahur, 2017) Balahur, Alexandra & Mohammad, Saif & van der Goot, Erik. (2017). Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2018, de Codalab Sitio web: <https://wt-public.emm4u.eu/wassa2017/cfp/cfp.pdf>
- (Barbosa, 2015) Barbosa, Liliana (2015). HACIA UN LEXICÓN UNIFICADO DE SENTIMIENTOS BASADO EN UNIDADES DE PROCESAMIENTO GRÁFICO (Tesis de Doctorado). Universidad Politécnica de Madrid, España
- (Becerra, 2017) Becerra, Martín. (2017). Analisis de sentimientos en Twitter: El bueno, el malo y el >:(. 2018, de EST, Concurso de Trabajos Estudiantiles Sitio web: <http://www.clei2017-46jaiio.sadio.org.ar/sites/default/files/Mem/EST/est-14.pdf>
- (Beigi, 2009) Beigi Ghazaleh & Hu Xia & Maciejewski Ross & Liu Huan. (2009). An Overview of Sentiment Analysis in Social Media and its Applications in Disaster Relief. 2018, de ONR Sitio web: <http://www.public.asu.edu/~gbeigi/files/BeigiSentimentChapter>.

pdf

- (Cambria, 2010) Cambria, Erik & Speer, Robert & Havasi, Catherine & Hussain, Amir. (2010). *SenticNet*. A Publicly Available Semantic Resource for Opinion Mining. 2018, de AAAI Fall Symposium Sitio web: <http://sentic.net/SenticNet.pdf>
- (Cambria, 2017) Cambria Erick & Das Dipankar & Bandyopadhyay Sivaji & Feraco Antonio. (2017). A Practical Guide to Sentiment Analysis. Canada: Springer. <http://sentic.net/practical-guide-to-sentiment-analysis.pdf>
- (Collomb, 2013) Collomb, Anaïs & Costea, Crina & Joyeux, Damien & Hasan, Omar & Brunie, Lionel. (2013). A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. 2018, de University of Lyon Sitio web: <http://liris.cnrs.fr/Documents/Liris-6508.pdf>
- (Contreras, 2016) Contreras, Fabián. (2016). INTRODUCCIÓN A MACHINE LEARNING. España: Sunqu https://www.zemsania.com/recursos-zemsania/whitepapers/DTS/Machine_learning.pdf
- (Conway, 2012) Conway Drew & Myles Jhon. (2012). Machine Learning for Hackers. United States of America: O'REILLY. http://pdf.th7.cn/down/files/1312/machine_learning_for_hackers.pdf
- (Cortez, 2002) Cortez, Augusto & Vega, Hugo & Pariona, Jaime. (2002). Procesamiento de lenguaje natural. 2018, de Universidad Nacional Mayor de San Marcos Sitio web: http://sisbib.unmsm.edu.pe/BibVirtual/Publicaciones/risi/2009_n2/v6n2/a06v6n2.pdf
- (Cho, 2013) Cho, Heeryon & Lee, Jong-Seok & Kim, Songkuk. (2013). Enhancing Lexicon-Based Review Classification by Merging and Revising Sentiment Dictionaries. 2018, de Technology, Yonsei University Sitio web: <https://pdfs.semanticscholar.org/243b/a445ab0d56196a6f26f106b48d767e9bf5c0.pdf>
- (Dodds, 2009) Dodds, Peter & Danforth, Christopher. (2009). Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. 2018, de ResearchGate Sitio web:

https://www.researchgate.net/publication/225520158_Measuring_the_Happiness_of_Large-Scale_Written_Expression_Songs_Blogs_and_Presidents?el=1_x_8&enrichId=rgreq-707fc57084615ab854916ae8eebabbeb-XXX&enrichSource=Y292ZXJQYWdlOzI2MTk1OTYxODtBUzoXNzUzNTg1MjM0ODYyMDhAMTQxODgyMDEzODYxMw==

- (Dubiau, 2013) Dubiau Luciana, M Ale Juan, Análisis de Sentimientos sobre un Corpus en Español: Experimentación con un Caso de Estudio. Facultad de Ingeniería, Universidad de Buenos Aires, Argentina, 2013 <http://42jaiio.sadio.org.ar/proceedings/simposios/Trabajos/ASAI/04.pdf>.
- (Dubiau, 2013) Dubiau, Luciana & Ale, Juan (2013). Procesamiento de Lenguaje Natural en Sistemas de Análisis de Sentimientos (Tesis de licenciatura). Universidad de Buenos Aires, Argentina.
- (Eres, 2009) Eres, Isabel & Almeida, Paulo. (2009). UN ESTUDIO SOBRE LOS CAMBIOS LINGÜÍSTICOS DEL ESPAÑOL ESCRITO EN LAS CHARLAS INFORMALES POR INTERNET. 2018, de USP Sao Paulo Sitio web: <http://www.scielo.br/pdf/tla/v48n1/11.pdf>
- (Esuli, 2010) Esuli, Andrea & Sebastiani, Fabrizio. (2010). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. 2018, de Instituto de Tecnología Italiana Sitio Web: <http://nmis.isti.cnr.it/sebastiani/Publications/LREC06.pdf>
- (Farooq, 2016) Farooq, Umar & Mansoor, Hasan & Nongailard, Antoine & Ouzrout, Yacine & Qadir, Muhammad. (2016). Negation Handling in Sentiment Analysis at Sentence Level. 2018, de Journal of Computers Sitio web: <http://www.jcomputers.us/vol12/jcp1205-11.pdf>
- (Fridja, 2011) Fridja, Nico & Mesquita Badja. (2011). An emotion perspective on emotion regulation. 2018, de Psychology Press Sitio web: <https://ppw.kuleuven.be/cscp/documents/mesquita/mesquita-frijda-2011-an-emotion-perspective-on.pdf>

- (Gallego, 2010) Gallego José. (2010). Comunidades virtuales y redes sociales. España: Wolters Kluwer. <https://www.kimera-mk.com/downloads/Comunidades-virtuales-y-redes-sociales.pdf>
- (Gelbukh, 2002) Gelbukh, Alexander. (2002). Procesamiento del lenguaje natural, 2018, de Sociedad Mexicana de Inteligencia Artificial Sitio web <https://www.gelbukh.com/CV/Publications/2010/Procesamiento%20de%20lenguaje%20natural%20y%20sus%20aplicaciones.pdf>.
- (Goncalves, 2013) Gonçalves, Pollyanna & Araújo, Matheus & Benevenuto, Fabrício & Cha, Meeyoung. (2013). Comparing and Combining Sentiment Analysis Methods. 2018, de UFMG Sitio web: <http://homepages.dcc.ufmg.br/~fabricio/download/cosn127-goncalves.pdf>
- (Gutiérrez, 2015) Gutiérrez, Guillermo (2015). Procesamiento del Lenguaje Natural Aplicado al Análisis del Sentimiento de Opiniones (Ingeniería en informática). Universidad de Cádiz, España.
- (Hernández, 2016) Hernández, Natalia (2016). METODOLOGÍA PARA EL DISEÑO Y CONSTRUCCIÓN DE UN LEXICÓN DE OPINIÓN, BASADO EN COMENTARIOS DE TWITTER APLICADO AL PROYECTO “OPINIONZOOM” (Tesis de licenciatura). Universidad de Chile, Chile.
- (Heffring, 14) Heffring Peter & Jenkins Erica. (Abril 2014). Inside Twitter: An In-Depth Look Inside the Twitter World. sysomos, 1, 21. <https://sysomos.com/wp-content/files/Inside-Twitter-BySysomos.pdf>
- (Hu, 2004) Hu, Mingqing & Liu, Bing. (2004). Mining and Summarizing Customer Reviews . 2018, de Department of Computer Science University of Illinois at Chicago Sitio web: <https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf>

- (Hurwitz, 2018) Hurwitz Judith & Kirsch Daniel. (2018). Machine Learning For Dummies. Hoboken NJ: John Wiley & Sons, Inc. http://www.cva.itesm.mx/biblioteca/pagina_con_formato_version_oct/apalibro.html
- (Hutto, 2014) Hutto C & Gilbert Eric. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. 2018, de Georgia Institute of Technology, Atlanta Sitio web: https://www.researchgate.net/publication/275828927_VADER_A_Parsimonious_Rule-based_Model_for_Sentiment_Analysis_of_Social_Media_Text
- (Islas, 2013) Islas Octavio & Ricuarte Paola. (2013). Investigar las redes sociales. Comunicación total en la sociedad de la ubicuidad. México D.F: Tecnológico del Monterrey. <http://editorialrazonypalabra.org/pdf/ryp/InvestigarRedesSociales.pdf>
- (Jones, 2014) Jones Steve. (2014). *Twitter and Society*. New York: Peter Lang. http://alexleavitt.com/media/2014_PeterLang_TwitterAndSociety.pdf
- (Kouloumpis, 2011) Kouloumpis, Efthymios & Wilson, Theresa & Moore, Jhoanna. (2011). *Twitter Sentiment Analysis: The Good the Bad and the OMG!*. 2018, de International AAAI Conference on Weblogs and Social Media Sitio web: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251>
- (Kumar, 2013) Kumar Shamanth & Morstatter Fred & Liu Huan. (2013). *Twitter Data Analytics*. India: Springer. <http://tweettracker.fulton.asu.edu/tda/TwitterDataAnalytics.pdf>
- (Liu, 2012) Liu Bing. (2012). *Sentiment Analysis and Opinion Mining*. California: Morgan & Claypool Publishers.
- (López, 2009) López, Erika (2009). *CARACTERIZACIÓN DE LAS FUNCIONES DE LOS EMOTICONES EN INTERACCIONES VIRTUALES DE CHAT ABIERTO* (Tesis de licenciatura). Pontificia Universidad Javeriana, Colombia.
- (Makrehchi, 2013) Makrehchi, Masoud & Shah, Sameena & Liao, Wenhui. (2013). *Stock Prediction Using Event-Based Sentiment Analysis*. 2018,

- de IEEE Sitio web:
<https://ieeexplore.ieee.org/document/6690034/>
- (Marsland, 2015) Marsland Stephen. (2015). MACHINE LEARNING An Algorithmic Perspective. Florida: CRC Press
https://doc.lagout.org/science/Artificial%20Intelligence/Machine%20learning/Machine%20Learning_%20An%20Algorithmic%20Perspective%20%282nd%20ed.%29%20%5BMarsland%202014-10-08%5D.pdf
- (Merodio, 2010) Merodio Juan. (2010). Marketing en redes sociales: mensajes de empresa para gente selectiva. España: creative commons.
http://www.apatgn.org/c/document_library/get_file?uuid=1fbd3475-a083-4e55-a28b-657cf2030576&groupId=10198
- (Mohri, 2012) Mohri, Mehryar & Rostamizadeh, Afshin & Talwalkar, Ameet. (2012). Foundations of Machine Learning. England: The MIT Press
[ftp://doc.nit.ac.ir/cee/jazayeri/MachineLearning/Mehryar_Mohri%202012/\(Mehryar_Mohri_Afshin_Rostamizadeh_Ameet_Talwalkar\(BookFi.org\).pdf](ftp://doc.nit.ac.ir/cee/jazayeri/MachineLearning/Mehryar_Mohri%202012/(Mehryar_Mohri_Afshin_Rostamizadeh_Ameet_Talwalkar(BookFi.org).pdf)
- (Montesinos, 2014) Montesinos, Lucas (2014). ANÁLISIS DE SENTIMIENTOS Y PREDICCIÓN DE EVENTOS EN TWITTER (Tesis de licenciatura). Universidad de Chile. Chile.
- (Musto, 2014) Musto, Cataldo & Semeraro, Giovanni & Polignano, Marco. (2014). A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts. 2018, de Department of Computer Science University of Bari Aldo Moro, Italy Sitio web:
<http://ceur-ws.org/Vol-1314/paper-06.pdf>
- (Nielssen, 2011) Nielssen, Finn. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. 2018, de DTU Informatics, Technical University of Denmark Sitio web:
<https://arxiv.org/pdf/1103.2903.pdf>
- (Nilsson, 1998) Nilsson Nils. (1998). INTRODUCTION TO MACHINE LEARNING AN EARLY DRAFT OF A PROPOSED TEXTBOOK. Stanford: Stanford University.
<http://robotics.stanford.edu/~nilsson/MLBOOK.pdf>

- (O'Reilly, 2012) O'Reilly Tim & Milstein Sarah. (2012). The Twitter Book. United States of America: O'Reilly Media, Inc. http://leotardi.ddns.info/download/The_Twitter_Book.pdf
- (Pak, 2015) Pak Alexander & Paroubek Patrick. (2015). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. 2018, de Université de Paris-Sud, Laboratoire LIMSIS-CNRS Sitio web: <https://pdfs.semanticscholar.org/ad8a/7f620a57478ff70045f97abc7aec9687ccbd.pdf>
- (Pallarés, 2010) Pallarés, Miguel. (2010). Emociones y sentimientos donde se forman y como se transforman. España: Marge books. <https://books.google.com.mx/books?id=vP5qbPFWp88C&prints=ec=frontcover&dq=Emociones&hl=es-419&sa=X&ved=0ahUKEWjGuqPvnrHbAhUFZawKHSGnA8MQ6AEILjAB#v=onepage&q=Emociones&f=false>
- (Ramos, 2009) Ramos Consuelo & Pedraza Diana. (2009). GUIA DE MANEJO DE REDES SOCIALES DE INTERNET. Perú: Organización de los Estados Americanos. <http://portal.oas.org/LinkClick.aspx?fileticket=RHldC3cS3Qw%3D&tabid=1483>
- (Read, 2005) Read Jonathon. (2005). Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. 2018, de University of Sussex Sitio web: http://delivery.acm.org/10.1145/1630000/1628969/p43-read.pdf?ip=177.247.103.76&id=1628969&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1530063796_47c64940328119e600138da422d9f7ae
- (Ríos, 2018) Ríos Raquel. (2018). ¿Qué ocurrió realmente? El escándalo de Facebook y Cambridge Analytica, 1, 37 https://www.reasonwhy.es/sites/default/files/libro_de_conocimiento_001-elescandalofacebook.pdf
- (Santiago, 2018) Santiago, Olga. (2018). Las emociones. 2018, de AECC Sitio web: <http://sauce.pntic.mec.es/falcon/emociones.pdf>
- (Sasank, 2016) Sasank, Ventaka & Reddy, Kamal & Panda, Ganapati. (2016). Sentiment Analysis of Twitter Data for Predicting Stock Market

- Movements. 2018, de International conference on Signal Processing, Communication, Power and Embedded System
Sitio web: <https://arxiv.org/pdf/1610.09225.pdf>
- (Shalev, 2014) Shalev Shai & David Shai. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press.
http://www.cva.itesm.mx/biblioteca/pagina_con_formato_version_oct/apalibro.html
- (Singh, 2018) Singh Gurjit & Gupta Gaurav. (2018). SENTIMENT ANALYSIS OF ENGLISH TWEETS USING DATA MINING: Data Mining, Sentiment Analysis. India: BookRix.
<https://books.google.com.mx/books?id=hSNTDwAAQBAJ&printsec=frontcover&dq=sentiment+analysis&hl=es-419&sa=X&ved=0ahUKEwjBwsys5vzaAhVihuAKHa0yBFU4ChDoAQg0MAI#v=onepage&q=sentiment%20analysis&f=false>
- (Strapparava, 2004) Strapparava, Carlo & Valitutti, Alessandro. (2004). WordNet-Affect: an Affective Extension of WordNet. 2018, de Istituto per la Ricerca Scientifica e Tecnologica Sitio web: <https://pdfs.semanticscholar.org/52b3/d9fcc8d0d3beb0357c598358f0435784d517.pdf>
- (Stone, 1963) Stone, Phillip & Hunt, Earl. (1963). A COMPUTER APPROACH TO CONTENT ANALYSIS: STUDIES USING THE GENERAL INQUIRER SYSTEM. 2018, de Harvard Sitio web: <https://www.computer.org/csdl/proceedings/afips/1963/5062/00/50620241.pdf>
- (Taboada, 2011) Taboada, Maite & Brooke, Julian & Tofiloski, Milan & Stede, Manfred. (2011). Lexicon-Based Methods for Sentiment Analysis. 2018, de Association for Computational Linguistics Sitio web: https://www.researchgate.net/publication/220355505_Lexicon-Based_Methods_for_Sentiment_Analysis
- (Tausczik, 2010) Tausczik, Yla & Pennebaker, James. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. 2018, de Journal of Language and Social Psychology Sitio web: <https://www.cs.cmu.edu/~ylataus/files/TausczikPennebaker2010.pdf>
- (Thelwall, 2013) Thelwall, Mike. (2013). Heart and Soul: Sentiment Strength Detection in the Social Web with SentiStrength. 2018, de

- Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton Sitio web: <http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf>
- (Urueña, 2010) Urueña Alberto. (2011). Las Redes Sociales en Internet. España: ontsi. http://www.ontsi.red.es/ontsi/sites/ontsi/files/redes_sociales-documento_0.pdf
- (Villena, 2017) Villena, Julio & García, Miguel & Martínez, Eugenio & Díaz, Manuel & García, Manuel. (2017). TASS-2017: Workshop on Semantic Analysis at SEPLN. 2018, de Sociedad Española para el Procesamiento del Lenguaje Natural Sitio web: <http://www.sepln.org/workshops/tass/2017/#organization>
- (Wang, 2012) Wang, Hao & Can, Dogan & Kazemzadeh, Abe & Bar, Francois & Narayanan. (2012). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. 2018, de ResearchGate Sitio web: https://www.researchgate.net/publication/262326668_A_System_for_Real-time_Twitter_Sentiment_Analysis_of_2012_US_Presidential_Election_Cycle?el=1_x_8&enrichId=rgreq-707fc57084615ab854916ae8eebabbeb-XXX&enrichSource=Y292ZXJQYWdIOzI2MTk1OTYxODtBUzoXNzUzNTg1MjM0ODYyMDhAMTQxODgyMDEzODYxMw==
- (Wiebe, 2005) Wiebe, Janyce & Wilson, Theresa & Cardie, Claire. (2005). Annotating Expressions of Opinions and Emotions in Language. 2018, de University of Pittsburgh Sitio web: <https://www.cs.cornell.edu/home/cardie/papers/lre05withappendix.pdf>
- (Zhang, 2011) Zhang, Lei & Ghosh, Riddhiman & Dekhil, Mohamed & Hsu, Meichun & Liu, Bing. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. 2018, de HP Sitio web: <http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf>

ANEXO A. Lexicones

A.1. Lexicón *VADER*

>:-)	-0.4	1.68523	[1, 2, 1, -2, -2, -1, -1, -3, -1, 2]
>:/	-1.6	0.8	[-1, -2, -1, -3, -1, -1, -1, -1, -2, -3]
>:o	-1.2	1.16619	[-3, -1, -2, 0, -2, -2, 0, -1, 1, -2]
>:p	1.0	0.7746	[-1, 1, 1, 2, 1, 2, 1, 1, 1, 1]
>:[-2.1	0.53852	[-2, -2, -2, -2, -3, -3, -2, -1, -2, -2]
>:\	-1.7	0.64031	[-1, -2, -1, -2, -2, -3, -1, -1, -2, -2]
>;(-2.9	0.7	[-3, -4, -3, -2, -2, -3, -3, -3, -2, -4]
>;)	0.1	1.04403	[-1, 1, 0, -1, 2, 0, -1, 1, 1, -1]
>_>^	2.1	0.9434	[2, 2, 1, 4, 3, 2, 1, 3, 1, 2]
@:	-2.1	0.9434	[-3, -2, -3, -1, -2, -4, -1, -2, -1, -2]
@>-->--	2.1	1.22066	[1, 1, 0, 2, 4, 2, 4, 2, 3, 2]
@}-;-'----	2.2	1.32665	[0, 1, 3, 2, 1, 4, 4, 1, 3, 3]
aas	2.5	0.80623	[2, 3, 3, 4, 1, 2, 3, 2, 2, 3]
aayf	2.7	0.78102	[2, 3, 2, 4, 3, 2, 2, 3, 4, 2]
afu	-2.9	0.83066	[-3, -3, -3, -3, -3, -1, -4, -4, -2, -3]
alol	2.8	0.74833	[2, 2, 2, 3, 3, 2, 3, 4, 4, 3]
ambw	2.9	0.7	[2, 3, 4, 2, 3, 2, 3, 3, 4, 3]
aml	3.4	0.66332	[4, 3, 2, 4, 3, 3, 4, 4, 3, 4]
atab	-1.9	1.22066	[-2, 0, -1, -2, -1, -1, -2, -4, -4, -2]
awol	-1.3	0.78102	[0, -1, -1, -1, -1, -1, -2, -2, -3, -1]
ayc	0.2	0.9798	[0, 1, -1, 1, 0, 1, 0, -1, 2, -1]
ayor	-1.2	0.6	[-1, -1, -2, -2, -1, -1, -1, 0, -2, -1]
aug-00	0.3	1.18743	[2, 0, -2, 0, 0, 1, -1, 2, 1, 0]
bfd	-2.7	0.78102	[-3, -2, -4, -2, -3, -2, -3, -2, -4, -2]

En el lexicón de *VADER* está estructurado por las palabras, la valencia de las palabras, la normalización de las calificaciones por los humanos y las calificaciones por los humanos mediante el AMT. Tomando en cuenta solamente la palabra y la valencia en el método de *VADER*.

Con el objetivo de obtener la valencia de cada palabra Hutto (2014) realiza el promedio de la calificación dada por el conjunto de 10 personas.

La evaluación de las palabras se obtuvo de la siguiente manera (2, 2, 2, 4, 2, 4, 3, 4, 2, 2) donde la suma de los valores sería 27 dividido entre 10 da como resultado que dicha palabra obtiene 2.7 por valencia.

A.2. Lexicón SentiStrength

X(-1
X-(-1
XD	1
XD	1
XO	-1
XP	-1
XP	1
^^	1
^o)	-1
x3?	1
xD	1
xP	-1
8C	-1
8c	-1
D	1
}:)	1
abandon	-2
abate	-2
abdicate*	-2
abhor*	-4
abject	-2
abnormal*	-2
abolish*	-2
abomina*	-3

El lexicón de SentiStrength se compone por las palabras y la valencia.

A.3. Lexicón SentiWordNet

_calibre _caliber -calibre -caliber	0
_may_organization	0
-karat_gold	0
a	0
a_battery	0
a_cappella	0
a_cappella_singing a_capella_singing	0
a_fortiori	0.25
a_la_carte	0
a_la_mode	0
a_level	0
a_posteriori	0.125
a_priori	0.25
a_trifle a_little a_bit	0
a-line	0
a-scan_ultrasonography	0
a-team	0
aa	0
aba	0
aback	0
abacus	0
abandonment	0.5
abarticulation	0
abasia	0.5

El lexicón de SentiWordNet se estructura por las palabras, sinónimos y valencias.

A.4. Lexicón *SenticNet*

a_little	-0.84	
a_little_hungry	0.757	
a_little_specific		0.06
a_lot	0.258	
a_lot_of_books	0.071	
a_lot_of_energy	0.766	
a_lot_of_fat	-0.95	
a_lot_of_flowers		0.041
a_lot_of_food	0.049	
a_lot_of_fun	0.835	
a_lot_of_information		0.104
a_lot_of_money	0.131	
a_lot_of_noise	-0.92	
a_lot_of_people	0.054	
a_lot_of_practice		0.875
a_lot_of_sex	0.858	
a_lot_of_space	0.851	
a_lot_of_stress	-0.10	
a_lot_of_study	-0.90	
a_lot_of_time	0.953	
a_lot_of_work	0.848	
aa_gill	-0.66	
aa_gun	-0.78	
abandon	-0.84	

El lexicón de *SenticNet* contiene las palabras y las valencias.

A.5. Lexicón Hu-Liu

amazing	1.4	
amazingly		2.3
ambitious		3.9
ambitiously		2.0
ameliorate		2.2
amenable		1.9
amenity	1.6	
amiability		1.7
amiably		1.4
amiable	2.3	
amicability		2.6
amicable		1.9
amicably		3.4
amity	3.6	
ample	3.3	
amply	3.7	
amuse	2.5	
amusing	3.7	
amusingly		3.3
angel	2.5	
angelic	3.6	
apotheosis		2.2
appeal	1.4	
appealing		1.7

El lexicón de Hu-Liu contiene las palabras y las valencias.

A.6. Lexicón SO-Cal

mega-success	5
priceless	5
exquisite	5
fascinating	5
phenomenal	5
top-notch	5
top-quality	5
top-rate	5
adorable	5
all-star	5
amazing	5
angelic	5
award-winning	5
awesome	5
best	5
best-drawn	5
best-handling	5
bestselling	5
breathtaking	5
breath-taking	5
brilliant	5
dazzling	5
decadent	5
delicious	5

El lexicón de SO-Cal se compone por las palabras y las valencias.

A.7. Lexicón ANEW

abduction	2.76
abortion	3.5
absurd	4.26
abundance	6.59
abuse	1.8
acceptance	7.98
accident	2.05
ace	6.88
ache	2.46
achievement	7.89
activate	5.46
addict	2.48
addicted	2.51
admired	7.74
adorable	7.81
adult	6.49
advantage	6.95
adventure	7.6
affection	8.39
afraid	2
aggressive	5.1
agility	6.46
agony	2.43
agreement	7.08

El lexicón de ANEW se estructura por las palabras y las valencias.

A.8. Lexicones combinados

wow	1.3	
wowed	1.0	
wowing	2.8	
wows	3.2	
yay	1.2	
youthful		2.6
zeal	3.3	
zenith	3.2	
zest	1.7	
zippy	3.9	
abnormal		-2.1
abolish	-2.1	
abominable		-1.9
abominably		-1.1
abominate		-2.5
abomination		-0.1
abort	-2.6	
aborted	-1.6	
aborts	-1.5	
abrade	-1.8	
abrasive		-2.3
abrupt	0.0	
abruptly		-0.1
abscond	-0.5	

La combinación de los lexicones se realizó uniendo los lexicones en un solo archivo de textos, debido a que el método *VADER* toma las palabras y la valencia del lexicon se puede realizar la evaluación de los Tweets, realizando la aproximación de combinación de lexicones vista en el estado del arte. Como se puede observar en la combinación presentada el primer lexicon termina en la palabra *zippy* y comienza con la palabra *abnormal*.

ANEXO B. Corpus

B.1. Corpus WASSA 2017 sin preprocesamiento

anger 10000 How the FUCK!!! Who the HELL!!!! moved my fridge!!! Should I knock the landlord door. angry mad 0.938
anger 10001 So my Indian Uber driver just called someone the N word. If I wasn't in a moving vehicle I would have jumped out disgusted 0.896
anger 10002 @DPD_UK I asked for my parcel to be delivered to a pick up store NOT my address!!! FUMING!!!! POOR!!!! customer service 0.896
anger 10003 so ef whichever butt wipe pulled the fire alarm in davis bc I was sound asleep #pissed #angry #upset #tired #sad #tired #hangry ##### 0.896
anger 10004 Don't join @BTCare they put the phone down on you, talk over you and are rude. Taking money out of my acc willynilly! #fuming 0.896
anger 10005 My blood is boiling 0.875
anger 10006 When you've still got a whole season of Wentworth to watch and a stupid cunt in work ruins it for us @ @_KirstyGA #raging #oldcunt 0.875
anger 10007 @bt_uk why does tracking show my equipment delivered, when it wasn't? Why is my service suddenly delayed? We've already 3 weeks. #fuming 0.875
anger 10008 @TeamShanny legit why i am so furious with him, people are such fucking idiots. 0.875
anger 10009 How is it suppose to work if you do that? Wtf dude? Thanks for pissing me off. 0.875
anger 10010 im so mad about power rangers. im incensed. im furious. 0.667
fear 20000 I feel like I am drowning. #depression #anxiety #failure #worthless 0.979
fear 20001 I get so nervous even thinking about talking to ***** I wanna die 0.979
fear 20002 I lost my blinders #panic 0.975
fear 20003 I feel like I am drowning. #depression #failure #worthless 0.938
fear 20004 This is the scariest American Horror Story out of all of them... I'm gonna have to watch in the daytime. #frightened 0.938
fear 20005 @mgcsartwork I nearly started crying and having a full on panic attack after tatinof bc of the crowds so I feel him 0.938
fear 20006 I have to finally tell my therapist about my sexuality ... last frontier ... not sure I can do it in the AM #fear #SingleGirlProblems 0.938
fear 20007 @laura221b I don't think I've ever moved so fast in a panic in all my life @ Gave me such a fright @ 0.938
fear 20008 My bus was in a car crash... I'm still shaking a bit... This week was an absolute horror and this was the icing on the cake... #terrible 0.938
fear 20009 My bus was in a car crash... I'm still shaking a bit... This week was an absolute horror and this was the icing on the cake... 0.920
fear 20010 procrastating is fun until im an hour away from the time its due and i still havent finished so i have a panic attack 0.920
joy 30000 Just got back from seeing @GaryDelaney in Burslem. AMAZING!! Face still hurts from laughing so much #hilarious 0.980
joy 30001 Oh dear an evening of absolute hilarity I don't think I have laughed so much in a long time! @ 0.958
joy 30002 Been waiting all week for this game ♥♥♥ #cheer #friday ♥ 0.940
joy 30003 @gardiner_love : Thank you so much, Gloria! You're so sweet, and thoughtful! You just made my day more joyful! I love you too! @♥ 0.938
joy 30004 I feel so blessed to work with the family that I nanny for ♥ nothing but love & appreciation, makes me smile. 0.938
joy 30005 Today I reached 1000 subscribers on YT!! , #goodday, #thankful 0.926
joy 30006 @Singaholic121 Good morning, love! Happy first day of fall. Let's make some awesome #autumnmemories #annabailey #laughter #smile 0.924
joy 30007 #BridgetJonesBaby is the best thing I've seen in ages! So funny, I've missed Bridget! #love #TeamMark 0.922
joy 30008 Just got back from seeing @GaryDelaney in Burslem. AMAZING!! Face still hurts from laughing so much 0.920
joy 30009 @IndyMN I thought the holidays could not get any more cheerful, and then I met you. #TheNiceBot 0.917
joy 30010 I'm just still . So happy .\nA blast 0.917
sadness 40000 Depression sucks! #depression 0.958
sadness 40001 Feeling worthless as always #depression 0.958
sadness 40002 Feeling worthless as always 0.958
sadness 40003 My #Fibromyalgia has been really bad lately which is not good for my mental state. I feel very overwhelmed #anxiety #bipolar #depression 0.946
sadness 40004 Im think ima lay in bed all day and sulk. Life is hitting me to hard rn 0.934
sadness 40005 So when i try I fail... and when I don't try... I still fail... #confused and #lost 0.917
sadness 40006 Extreme sadness 0.917
sadness 40007 my life in one word is depressing 0.917
sadness 40008 Panic attacks are the worst. Feeling really sick and still shaking. I should be a sleep. #anxiety #depression 0.917
sadness 40009 Feel so grim + ugly atm 0.896
sadness 40010 Honestly depression is kicking my ass lately.. @ 0.896

B.2. Corpus WASSA 2017 con preprocesamiento

anger 10000 How the FUCK!!! Who the HELL!!!! moved my fridge!!! Should I knock the landlord door. ANGRY!!! mad 0.938
anger 10001 So my Indian Uber driver just called someone the N word. If I FEAR!!! in a moving vehicle I would have jumped out disgusted 0.896
anger 10002 @DPD_UK I asked for my parcel to be delivered to a pick up store NOT my address!!! FUMING!!!!!! POOR!!!! customer service 0.896
anger 10003 so ef whichever butt wipe pulled the fire alarm in davis bc I was sound asleep pissed ANGRY!!!! upset tired sad tired HANGRY!!!! 0.896
anger 10004 :(join @BTCare they put the phone down on you, talk over you and are rude. Taking money out of my acc willynilly! FUMING!!!! 0.896
anger 10005 My blood is BOILING!!!! :(0.875
anger 10006 When you've still got a whole season of Wentworth to watch and a stupid cunt in work ruins it for us @ @_KirstyGA raging oldcunt 0.875
anger 10007 @bt_uk why does tracking show my equipment delivered, when it FEAR!!!!? Why is my service suddenly delayed? We've already 3 weeks. FUMING!!!! 0.875
anger 10008 @TeamShanny legit why i am so FURIOUS!!! with him, people are such fucking idiots. 0.875
anger 10009 How is it suppose to work if you do that? Wtf dude? Thanks for pissing me off. 0.875
anger 10010 im so mad about power rangers. im incensed. im FURIOUS!!!!. 0.667

fear 20000 I feel HATE!!! I am drowning. depression anxiety failure worthless 0.979
 fear 20001 I get so nervous even thinking about talking to ***** I wanna die 0.979
 fear 20002 I lost my blinders panic 0.975
 fear 20003 I feel HATE!!! I am drowning. depression failure worthless 0.938
 fear 20004 This is the scariest American Horror Story out of all of them... I'm gonna have to watch in the daytime. frightened 0.938
 fear 20005 @mgcsartwork I nearly started crying and having a full on panic attack after tatinof bc of the crowds so I feel him 0.938
 fear 20006 I have to finally tell my therapist about my sexuality ... last frontier ... :(sure I can do it in the AM fear SingleGirlProblems 0.938
 fear 20007 @laura221b I >:(think I've ever moved so fast in a panic in all my life @ Gave me such a fright @ 0.938
 fear 20008 My bus was in a car crash... I'm still shaking a bit... This week was an absolute horror and this was the icing on the cake... terrible 0.938
 fear 20009 My bus was in a car crash... I'm still shaking a bit... This week was an absolute horror and this was the icing on the cake... 0.920
 fear 20010 procrastinating is spite until im an hour away from the time its due and i still havent finished so i have a panic attack 0.920
 joy 30000 Just got back from seeing @GaryDelaney in Burslem. AMAZING!! Face still hurts from laughing so much hilarious 0.980
 joy 30001 Oh dear an evening of absolute hilarity I sete think I have laughed so much in a long time! @ 0.958
 joy 30002 Been waiting all week for this game ♥♥♥ cheer friday ♥0.940
 joy 30003 @gardiner_love : Thank you so much, Gloria! You're so sweet, and thoughtful! You just made my day more joyful! I LOVE!!! you too! @♥ 0.938
 joy 30004 I feel so blessed to work with the family that I nanny for ♥ nothing but LOVE!!! & appreciation, makes me smile. 0.938
 joy 30005 Today I reached 1000 subscribers on YT!! , goodday, thankful 0.926
 joy 30006 @Singaholic121 Good morning, LOVE!!!! Happy first day of fall. Let's make some awesome autumnmemories annabailey laughter smile 0.924
 joy 30007 BridgetJonesBaby is the best thing I've seen in ages! So funny, I've missed Bridget! LOVE!!! TeamMark 0.922
 joy 30008 Just got back from seeing @GaryDelaney in Burslem. AMAZING!! Face still hurts from laughing so much 0.920
 joy 30009 @IndyMN I thought the holidays could not get any more cheerful, and then I met you. TheNiceBot 0.917
 joy 30010 I'm just still . So happy .\nA blast 0.917
 sadness 40000 Depression sucks! depression 0.958
 sadness 40001 Feeling worthless as always depression 0.958
 sadness 40002 Feeling worthless as always 0.958
 sadness 40003 My Fibromyalgia has been really bad lately which is :(good for my mental state. I feel very overwhelmed anxiety bipolar depression 0.946
 sadness 40004 Im think ima lay in bed all day and sulk. Life is hitting me to hard rn 0.934
 sadness 40005 So when I try I fail... and when I >:(try... I still fail... confused and lost 0.917
 sadness 40006 Extreme sadness 0.917
 sadness 40007 my life in one word is depressing 0.917
 sadness 40008 Panic attacks are the worst. Feeling really sick and still shaking. I should be a sleep. anxiety depression 0.917
 sadness 40009 Feel so grim + ugly atm 0.896
 sadness 40010 Honestly depression is kicking my ass lately.. @ 0.896

ANEXO C. Lista de comandos utilizados y scripts

C.1. Comandos.

Comando	Descripción
pip install virtualenv	Instalación de entorno virtual para proyectos en Python
pip install virtualenvwrapper-win	Mantenimiento del entorno virtual de Python
mkvirtualenv QlikSenseAAI	Creación de carpeta virtual en Python QlikSenseAAI
pip install grpcio	Instalación de librería grpcio, realiza peticiones entre cliente y servidor
pip install requests	Instalación de librería requests, realiza peticiones en entre cliente y servidor
pip install VADERSentiment	Instalación del método VADER.
workon QlikSenseAAI	Inicia la carpeta virtual donde se encuentra el método VADER.
python __main__.py	Librería de conexión entre VADER Y QlikSense.

C.2. Scripts

En la siguiente ilustración se muestra el script desarrollado en QlikSense donde se realiza la carga de las opiniones del corpus de Twitter de WASSA 2017.

```
1 WASSA_2017:
2 LOAD
3     Emotion,
4     Id,
5     Tweet,
6     Score //CAMPOS DEL CORPUS DE TWITTER DE WASSA 2017
7 FROM [lib://Archivos (win-rjcdromf96s_administrator)/TXT\TwitterWASSA2017Unicode.txt]
8 (txt, unicode, embedded labels, delimiter is '\t', msq);
9
10 // TENSOR
11 ArticleSentiment:
12 LOAD
13     *,
14     If([Article Title Sentiment - Compound]<=-.66,'Very Negative',
15     If([Article Title Sentiment - Compound]<=-.33,'Negative',
16     If([Article Title Sentiment - Compound]<=-0,'Somewhat Negative',
17     If([Article Title Sentiment - Compound]<=.33,'Somewhat Positive',
18     If([Article Title Sentiment - Compound]<=.66,'Positive','Very Positive'))))
19 AS "Article Title Sentiment - Compound Buckets" //POLARIDADES NORMALIZADAS DE -1 A 1
20 ;
21 LOAD
22     *,
23     TextBetween("Article Title Sentiment",'neg: ','|') AS "Article Title Sentiment - Negative",
24     TextBetween("Article Title Sentiment",'compound: ','|') AS "Article Title Sentiment - Compound",
25     TextBetween("Article Title Sentiment",'pos: ','|') AS "Article Title Sentiment - Positive",
26     TextBetween("Article Title Sentiment",'neu: ','|') AS "Article Title Sentiment - Neutral"
27 ;
28 LOAD
29     Field1 AS Id,
30     Field2 AS "Article Title Sentiment"
31 EXTENSION PythonSentiment.SentimentScript (WASSA_2017("Id"."Tweet")):
```

Se puede apreciar que la carga de datos se realiza de las líneas 1 a la 8 donde se coloca la ruta del archivo del corpus y los campos de los datos del corpus.

De la línea 14 a la 31 se realiza la clasificación de polaridades de las opiniones del corpus donde se analizan desde la librería “EXTENSION PythonSentiment.SentimentScript(WASSA_2017{"Id","Tweet"});”. Una vez obtenidas las polaridades se van analizando en las líneas 14 a la 19, se realizó dicha aproximación de polaridades para obtener una mejor clasificación de estas.

Con el objetivo de realizar la igualdad entre polaridades y emociones descrita en capítulos anteriores se realizó los siguientes scripts.

```
Count({<[Article Title Sentiment - Compound]={">0"}, Emotion = {'joy'}}>Id)

Count({<[Article Title Sentiment - Compound]={"<=0"}, Emotion = {'anger','fear','sadness'}}>Id)
```

Se realiza el conteo de opiniones de Twitter del corpus cargado, donde la polaridad de la opinión sea positiva (mayor a 0) y donde la emoción del corpus de WASSA este etiquetada como felicidad ('joy'). De igual forma para el conteo de las polaridades negativas y neutras (menor o igual a 0) con respecto a las emociones del corpus de WASSA etiquetadas como furia, miedo y tristeza ('anger, fear & sadness').

C.3. Script F-Measure

Para realiza el cálculo se crearon cuatro variables *a*, *b*, *c* y *d* (mencionando dichas clasificaciones en capítulo anteriores).

Donde *a* se da por el total de opiniones positivas correctas.

```
Count({<[Article Title Sentiment - Compound]={">0"}, Emotion = {'joy'}}>Id)
```

Donde *b* se da por el total de opiniones etiquetadas correctas del corpus de WASSA menos las etiquetadas por el método *VADER*. Donde el resultado de la resta son las opiniones negativas clasificadas como positivas.

```
Count({<Emotion={'anger','fear','sadness'}>>Id) - Count({<[Article Title Sentiment - Compound]={"<=0"}, Emotion = {'anger','fear','sadness'}>>Id)
```

Donde c se da por el total de opiniones etiquetadas correctas del corpus de WASSA menos las etiquetadas por el método *VADER*. Donde el resultado de la resta son las opiniones positivas clasificadas como negativas.

```
Count({<Emotion={'joy'}>>Id) - Count({<[Article Title Sentiment - Compound]={">0"}, Emotion = {'joy'}>>Id)
```

Donde d se da por el total de opiniones negativas correctas.

```
Count({<[Article Title Sentiment - Compound]={"<=0"}, Emotion = {'anger','fear','sadness'}>>Id)
```

De acuerdo con las fórmulas de precisión, exhaustividad y medida F, se realizaron de la siguiente forma.

Precisión (guardado en una variable P).

$$= \frac{a}{(a + b)}$$

Exhaustividad (guardado en una variable R).

$$= \frac{a}{(a + c)}$$

Medida F.

$$= \frac{2 * (P * R)}{(P + R)}$$