UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE INGENIERÍA DE SISTEMAS

ANÁLISIS DE SENTIMIENTOS PARA TEXTOS CORTOS EN ESPAÑOL, UNA REVISIÓN DEL ESTADO DEL ARTE

Trabajo de titulación previo a la obtención del

Título de Ingeniera de Sistemas

AUTORA: Jessica Gabriela Valladares Cedillo

TUTOR: Julio Ricardo Proaño Orellana

Quito – Ecuador

2022

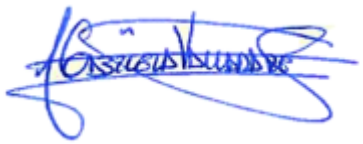
CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Yo, Jessica Gabriela Valladares Cedillo con documento de identificación N° 1722555800, manifiesto que:

Soy la autora y responsable del presente trabajo; y, autorizo a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 08 de marzo del 2022

Atentamente,



Jessica Gabriela Valladares Cedillo

1722555800

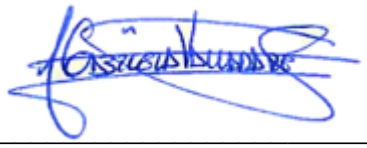
CERTIFICADO DE SESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Yo, Jessica Gabriela Valladares Cedillo, con número de cédula No. 1722555800, expreso mi voluntad y por medio del presente cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud que soy la autora del Artículo Académico: "Análisis de Sentimientos para Textos Cortos en Español, una Revisión del Estado del Arte.", el cual ha sido desarrollado para optar por el título de Ingeniera de Sistemas, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 08 de marzo del 2022

Atentamente,



Jessica Gabriela Valladares Cedillo

1722555800

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN.

Yo Julio Ricardo Proaño Orellana con C.I.: 0103909412, docente de la Universidad Politécnica Salesiana, declaro que bajo mi dirección y asesoría fue desarrollado el Trabajo de Titulación: ANÁLISIS DE SENTIMIENTOS DE TEXTOS CORTOS EN ESPAÑOL, UNA REVISIÓN DEL ESTADO DEL ARTE, realizado por Jessica Gabriela Valladares Cedillo, obteniendo como resultado final el trabajo de titulación bajo la opción de Artículo Académico con lo cual cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 08 de marzo del 2022

Ing. Julio Ricardo Proaño Orellana, PhD.

0103909412

DEDICATORIA

A mi madre, quien a pesar de todas las dificultades no ha permitido que me dé por vencida, y con su ejemplo me inspiró a seguir adelante para cumplir mi sueño ahora hecho realidad Mami tu bendición a diario y a lo largo de mi vida me protege y me lleva por el camino correcto. Gracias Licenciada Silvia Cedillo, porque he llegado hasta aquí por ti, te amo.

A mi amada hija, razón y motor de vida que me impulsa a superarme día a día, principal cimiento de formación y construcción en todos los ámbitos de mi vida.

A mi hermana, mi mejor amiga, mi futura colega quien me sostiene y me alienta cuando estoy a punto de caer, mi orgullo, mi inspiración y mi eterna compañera en el recorrido de cada vuelta al Sol.

Este título es para ustedes princesas de mi vida.

AGRADECIMIENTOS

A Dios con mucho amor y gratitud por acompañarme durante mi profesión, eterno forjador de esperanza lucha y valentía por haberme permitido gozar de salud para terminar el proyecto de mi tesis profesional.

A Alexis, mi esposo por su apoyo incondicional, por su comprensión y compañía para subir un peldaño más.

A Boris, mi mejor amigo quien entre regañada y risa me animó en aquellas ocasiones de duro caminar.

A mis tíos, y tías a toda mi familia y en especial a Netito y Blanquita mis abuelitos pilares fundamentales desde mi niñez, haciendo de mí la niña de sus ojos, regalándome su perfecto amor y cuidado en cada paso que yo he dado.

Hoy he cumplido una meta más de una lista interminable de metas, gracias a todos quienes me acompañaron en este proceso.

Jessica Gabriela Valladares Cedillo

ANÁLISIS DE SENTIMIENTOS DE TEXTOS CORTOS EN ESPAÑOL, UNA REVISIÓN DEL ESTADO DEL ARTE

SENTIMENT ANALYSIS FOR SHORT TEXT IN SPANISH, A REVIEW OF STATE OF THE ART

Jessica Gabriela Valladares Cedillo[1], Julio Ricardo Proaño Orellana[2]

RESUMEN

Actualmente las redes sociales son el medio de comunicación más utilizado por los usuarios en general. El análisis automático de las opiniones y comentarios emitidos en temas de interés como, por ejemplo: ciencia, tecnología, política, etc., requieren una revisión exhaustiva que gracias al análisis de sentimientos se logra determinar lo que los usuarios quieren expresar. En la actualidad existen una gran cantidad de herramientas que permiten realizar este análisis de sentimientos para textos cortos. Sin embargo, la mayoría se enfoca en el idioma inglés.

Este artículo tiene como objetivo el estudio de herramientas y Corpus utilizados para el análisis de sentimientos de textos en español mediante un estudio del estado del arte. Entre los resultados más importantes se encontró que el corpus más utilizado es el TASS, Además se realizó una comparación entre los métodos utilizados, entre ellos se pueden destacar SentiWordNet, Bayes, iSol, siendo el más eficiente SVM.

Palabras clave: Procesamiento de lenguaje natural (PLN), minería de textos, análisis de sentimientos, minería de opiniones.

ABSTRACT

In general, social networks are currently used by users for communication, comments, and opinions. So, the automatic analysis of them on topics of interest requires an exhaustive review by machines. With sentiment analysis, it is possible to determine what users want to express. Currently, many tools allow this sentiment analysis to be carried out for short texts. However, most of them are focused on the English language.

This article aims to study the tools and Corpus used to analyze sentiments of texts in Spanish through state of art. Among the most important results, it was found that the most used Corpus is the TASS. In addition, a comparison was made between the methods used, including SentiWordNet, Bayes, iSol, the most efficient being SVM.

Keywords: Natural Language Processing (NLP), Text Mining, Sentiment Analysis, Opinion Mining.

1 Estudiante de Ingeniería de Sistemas-Universidad Politécnica Salesiana- Sede Quito- Correo institucional: jvalladares@est.ups.edu.ec

[2] PhD. Julio Proaño Orellana Docente Titular Principal Nivel 1-Universidad Politécnica Salesiana- Sede Quito – Correo institucional:jproanoo@ups.edu.ec

1. Introducción

El uso de redes sociales en los últimos años se ha incrementado de tal manera que los usuarios interactúan prácticamente en tiempo real, expresando sus emociones, sentimientos, comentarios, opiniones, con respecto a cualquier tipo de situación o publicación. Existen páginas dedicadas a la recolección de opiniones como blogs y foros, aunque la mayoría de los CORPUS recolectan palabras directamente de Twitter.

El análisis de sentimientos se enfoca en determinar de manera automática la polaridad del texto, es decir, si este expresa un sentimiento negativo o positivo o si a su vez no expresa ningún sentimiento (neutro), lo cual para las pequeñas o grandes empresas es de mucha ayuda puesto que el análisis de sentimientos colabora de manera significativa en el proceso de toma de decisiones. Para realizar este análisis existen dos enfoques principales que son el enfoque LÉXICO o semántico y el de aprendizaje automático (ML). Cada uno se compone de distintas técnicas que permiten clasificar palabra dándoles una polaridad (positiva, negativa, neutra) [11].

1.1 Proceso para minería de texto

Para realizar el análisis de sentimientos el texto debe pasar por un preentrenamiento, proceso descrito a continuación.

1) Normalización del texto

Consiste en normalizar el texto en sólo minúsculas.

2) Tokenización

Consiste en separar cada palabra de la oración individualmente, incluye los signos como puntos y comas, creando así el diccionario de datos o listado de palabras.

3) Limpieza de información (Normalización)

Consiste en eliminar los signos de puntuación como los puntos, comas, tildes, interrogación, admiración, entre otros.

4) Filtrado de palabras vacías (stopwords)

Consiste en eliminar todas las palabras que no tengo un significado relevante, entre estas los artículos como “a”, “de”, “la”, “el”, entro otros

5) Lematización

Consiste en contar las palabras que se repiten en el diccionario de datos a partir de su raíz lingüística, lo primero que hace es devolver el verbo a su tiempo natural, es decir, convierte todas las palabras que en contexto signifiquen lo mismo, por ejemplo: si se desea saber cuántas veces aparece la palabra “cerrado” debemos contar la cantidad de veces que se encuentran las palabras “cerrar”, “cierra”, “cerrado”; está transformación de la palabra se define como DERIVACIÓN o STEMMING.[25]

1.2 Análisis del texto

Existen tres niveles de análisis de textos:

1. Análisis a nivel de documentos: Determina si un documento expresa sentimientos positivos, negativos o neutros.

2. Análisis a nivel de frases: Determina si una frase está expresando una opinión (frase subjetiva) o describe una información concreta (frase objetiva).

3. Análisis a nivel de entidad o palabra: Determina el tema sobre el cual se está opinando según las palabras. [25]

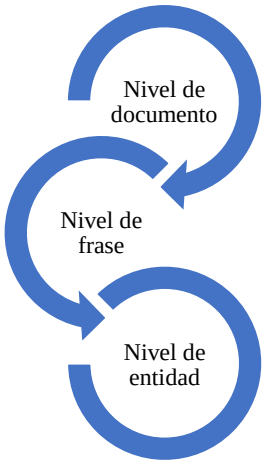


Figura 1. Niveles de análisis del texto.

1.3 Proceso de validación

Para realizar la validación en [5] se propone generar una matriz de confusión que observamos en la Tabla 1, en la que se colocan predicciones vs realidad; que permite verificar las palabras que han sido clasificadas de manera correcta, mediante las fórmula de precisión, recuperación y F-medida, en la que, precisión determina los errores cometidos en la sustitución e inserción, dividiendo el número de palabras correctamente identificadas como verdaderas positivas (VP) para la suma de éstas más las identificadas como falsas positivas (FP), es decir, aquellas palabras que se clasificaron como positivas pero resultaron ser negativas.

$$Precisión(P) = \frac{VP}{(VP + FP)} \qquad (1)$$

La recuperación determina cuantas palabras fueron predichas correctamente como positivas (VP) divido para la suma de estas más las falsas negativas (FN), es decir, aquellas que se clasifican como negativas, pero resultaron ser positivas, esta comparativa es llamada matriz de confusión (ver Tabla 1)

$$Recuperación(R) = \frac{VP}{(VP + FN)} \qquad (2)$$

Tabla 1. Matriz de confusión

	Positivo	**Negativo**
Positivo	Verdadero positivo	Falso negativo
Negativo	Falso negativo	Verdadero negativo

F-medida combina precisión y recuperación para determinar si existe un equilibrio de mediciones en el aprendizaje o análisis de las palabras.

$$F_m = \frac{2 * P * R}{(P + R)} \qquad (3)$$

1.4 Diccionarios, técnicas y enfoques; conceptos

Corpus: Diccionario de palabras recolectadas de un conjunto de opiniones con fines investigativos.

SEL: Diccionario de palabras conformada por alrededor de 2036 palabras, obtenidas principalmente de tweets compartidos por los usuarios, que determina si una palabra utilizada con frecuencia expresa uno de los 6 léxicos (P+, P, NEU, N, N-) considerados en SEL mediante el factor de uso efectivo (FPA).

LWIC: Diccionario compuesto por alrededor de 12656 palabras clasificadas entre los 4 léxicos considerados (P, NEU, N, Ninguno), cuenta con 464 categorías, destacándose entre ellas la categoría de procesamiento afectivo en la que determina si una palabra expresa un sentimiento positivo o negativo.

Raíz de una palabra es la parte que no cambia de ésta y permite formar otras palabras con significados acordes, ejemplo: FLOR es la raíz de la palabra y podemos generar: FLOReria, FLORecer, FLORista, etc.

Estilometría: Disciplina que permite un análisis automatizado del corpus.

2. Metodología

2.1 Descripción de la metodología

Para la realización de este estudio se hace una revisión estándar de literatura tomando como base la revisión de literatura sistemática, que consiste en la selección de artículos de repositorios que contengan la información que permita responder a las preguntas de investigación, para lo cual se siguen pasos divididos en 3 fases:

1) Búsqueda
2) Selección
3) Resultados.

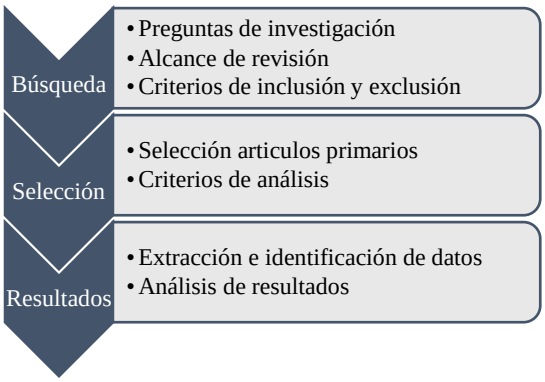


Figura 2. Fases de investigación

2.1.1 Fase 1

A continuación, se definen los términos descritos en la tabla 3 que permiten realizar la búsqueda con la ayuda de expresiones booleanas AND y NOT que definen la cadena de búsqueda, por ejemplo, Sentiment Analysis "AND" Spanish, Mining opinión "NOT" english.

Tabla 2. Palabras clave

Español	**Inglés**	**Significado/objetivo**
Análisis de sentimientos	Sentiment Analysis	Determinación del tono emocional de un conjunto de palabras o frases.
Procesamiento natural de lenguaje (PNL)	Natural language processing	Proceso de las comunicaciones humano-máquina.
Minería de opinión	Opinion mining	Interpreta el contexto o emoción de una frase.
Minería de textos	Text mining	Determina si existe un nuevo significado en el texto.

Preguntas de investigación

La definición de preguntas de investigación son una guía base para la recopilación de artículos a estudiar, se definen preguntas de revisión de la literatura (SLRP).

- ✓ ***SLRP1:*** ¿Cómo se construyen los corpus o data set que permitan realizar el SA y cuál es el más utilizado?
- ✓ ***SLRP2:*** ¿Cuál es el enfoque más utilizado para SA?
- ✓ ***SLRP3:*** ¿Cuáles son las técnicas que más se utilizan para el SA y bajo que enfoque?
- ✓ ***SLRP4:*** ¿Qué técnica y enfoque demuestra mayor eficacia?

Alcance de revisión

Existe gran cantidad de información en cuanto a análisis de sentimientos, sin embargo, al tratarse del idioma español esta información es aún escasa, se realizó el estudio del estado del arte de artículos que brindaron información de análisis de sentimientos de textos en español encontrados en repositorios de acceso abierto y de suscripción que permiten realizar una buena investigación que aporte a trabajos futuros.

Se realizó una revisión de literatura estándar tomando como base la revisión de literatura sistemática para complementar la investigación.

Criterios de inclusión y exclusión

Incluir artículos cuya investigación se base en el análisis de sentimientos en el idioma español, desde el año 2009 hasta el año 2021.

2.1.2 Fase 2

Se realizó la búsqueda en 5 repositorios (ver Tabla 4) que proporcionaron artículos que garantizan contar con información suficiente para dar respuesta a las preguntas establecidas en el SLRP.

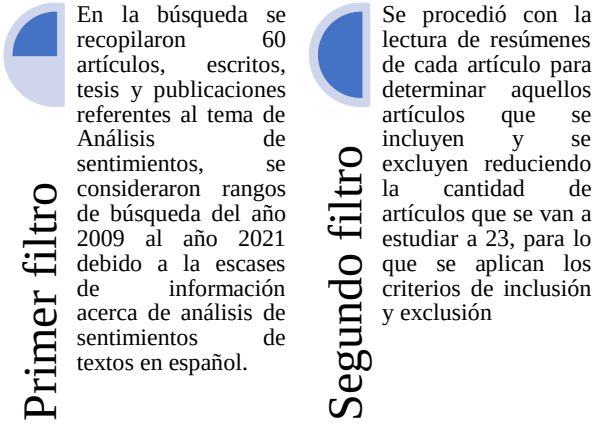


Figura 3. Filtros de búsuqeda y selección de artículos.

Tabla 3. Repositorios de búsqueda

	Tipo artículo
Scopus	Revistas
Research Gate	Revistas y conferencias
Emerald insgith	Revistas, tesis
Web of science	Revistas
Science Direct	Revistas

2.1.3 Fase 3

Extracción de datos

Se generó una matriz de referencias (ver tabla 10 en anexo 1) en la cual se detalló el año del artículo, autor, ideas principales, enfoques y técnicas utilizados por los autores.

En la tabla 4 se detalla los enfoques y técnicas utilizadas encontradas.

Tabla 4. Enfoques y técnicas

ENFOQUE	TÉCNICA
Machine Learning	CNN Naive Bayes SePLN SVM Bayesnet TwilBert (Bert) Isol Esol
Léxico	Tass Sentiword Word2vec SGS Inegi

Identificación de datos

Se enumeran los artículos utilizados para la revisión de la siguiente manera: [1],[2],[3],[4],[5],[6],[7],[9],[10],[11],[12], [13],[14],[15],[16],[17],[18],[19],[20],[21], [22],[23],[24].

En la Tabla 5 se listan los artículos seleccionados para la revisión del estado del arte, en dónde el id identifica la referencia del artículo, el data set indica el diccionario que fue utilizado en dicho artículo, se describen también el año y la técnica utilizada por el autor para realizar el análisis de sentimientos.

Tabla 5. Artículos relacionados

Artículos relacionados			
#Id	**Año**	**Dataset**	**Técnica**
[1]	2014	CORPUS	Naive Bayes and SVM
[2]	2016	Twitter	ANEW
[3]	2009	Corpus	SO Calculation - SVM
[4]	2017	Twitter	INEGI and TASS'15
[5]	2017	Corpus	LWIC
[6]	2018	Twitter	SEPLN
[7]	2016	POS(Movie and product)	Machine Learning (ML)
[9]	2018	Twitter	SMO - BayesNet-J48
[10]	2017	Twitter	CNN y word2vec
[11]	2016	Twitter	CNN / SVM
[12]	2013	CORPUS	SentiWordNet
[13]	2020	Twitter	TWiLBERT (BERT)
[14]	2019	Twitter	TASS
[15]	2014	Twitter	TASS
[16]	2015	WEB	SentiWordNet iSol eSol
[17]	2020	Corpus	-
[18]	2020	Corpus	Machine Learning
[19]	2019	Twitter	iSol
[20]	2021	Corpus	TwilBert (Bert)
[21]	2013	Corpus	SentiText
[22]	2020	Corpus	Naive Bayes
[23]	2017	Corpus	LIWC & Sel
[24]	2016	Twitter	

Análisis de resultados

En este último paso se responden las preguntas planteadas para la revisión de la literatura.

Se responde cada pregunta de manera cualitativa o cuantitativa según ésta se encuentre planteada.

3. Resultados

De acuerdo con lo investigado, existe una gran cantidad de CORPUS utilizados para realizar el análisis de sentimientos, la mayoría de los corpus están compuestos por publicaciones de twitter, otras se obtienen de foros, blogs y páginas dedicadas a recolección de opiniones de distintas áreas como turismo, restaurantes, finanzas, entre otros [17]; algunos corpus que existen son COAH, COAR, TASS, entre otros, siendo el más grande TASS ya que contiene más de 70000 tweets (ver Tabla 6), creado en el año 2012 con su última versión en el 2020.

Tabla 6. Corpus

Corpus	Sector	Recurso	Cant. Opiniones
HOpinion	Turismo	TripAdvisor	17934
Grueso	Turismo	TripAdvisor	2202
COAH	Turismo	TripAdvisor	1816
Coste	Publicaciones informales	Twitter	34634
COPOS	Salud	Opiniones web	743
COAR	Restaurantes	TripAdvisor	2202
TASS	Tweets (TV- Política, entre otros)	Twitter	Más de 70000 tweets clasificados en distintas áreas.

Cuenta con dos tipos de evaluaciones basada en 6 y 4 clasificaciones [11] a las cuales se le han otorgado rangos [5] (ver tabla 7).

Tabla 7. Clasificación de polaridad

CLASIFICACIÓN DE POLARIDAD		
RANGO	**SENTIMIENTO**	**SIMBOLOGÍA**
5	Muy positivo	P+
4	Positivo	P
3	Neutro	NEU
2	Negativo	N
1	Muy negativo	N+
	Ninguno	NONE / NINGUNO

En [5],[6],[11],[24],[20],[13],[14],[15], [21] los autores analizan el corpus TASS utilizando los léxicos SEL y LIWC; dos diccionarios que están comprendidos por alrededor de 2036 y 12656 palabras y raíces clasificadas en éstas 4 dimensiones:

1. Proceso lingüístico estándar.
2. Proceso psicológico.
3. Relatividad.
4. Preocupaciones personales.

Al no contar con características lingüísticas completas utilizan variables estilométricas, las cuales permiten cuantificar la frecuencia con la que la palabra es utilizada, siendo LIWC el más eficiente puesto que devuelve indicadores que permiten clasificar palabras según su polaridad (Positiva, negativa o neutra) [25].

Además, se comparan los algoritmos de aprendizaje automático (ML) SVM, J48 y NaiveBayes [22], entre los cuales destaca SVM ya que generaliza espacios de características de alta dimensión y facilita la categorización de texto al eliminar la necesidad de uso de funciones y además de ser un algoritmo robusto.

Además de las palabras, los llamados “emoticonos” [11]. Son también tomados en cuenta para el análisis de sentimientos, puesto que actualmente la mayoría de las personas utilizan estos símbolos para expresar o comunicar un sentimiento, los emoticonos son también clasificados según su polaridad en positivos, negativos o neutros, siguiente un conjunto de reglas [11] que permite determinar lo que un usuario quiere expresar (ver tabla 8).

Tabla 8. Regla de emoticonos

REGLA DE EMOTICONOS			
POS	**NEG**	**REGLA**	**CLASIFICACION**
:) **(:** **;)** **:-)** **(-:** **:D**	**:(** **):** **;(** **:-(** **)-:** **D:**	Sólo positivos	Positivo
		Más positivos que negativos	Positivo
:D **:P** **:-P**	**D-:** **:´-(** **)-´:**	Igual positivos que negativos	Neutro
		Más negativos que positivos	Negativos
		Sólo negativos	Negativos

El modelo encargado de la clasificación de emoticonos y palabras es llamado Cardellino, un modelo pre entrenado conformado por más de 1 millón de palabras, al ser comparado con otro modelo compuesto por 68.000 tweets resulta no arrojar resultados precisos, para

realizar esta comparación se utiliza el algoritmo de modelo de red neural 'word2vec' que maneja vectores que permiten determinar la polaridad de la palabra, mientras que para cumplir este mismo objetivo el modelo TwilBert [13] y [20] utiliza métricas que se encargan de eliminar redundancia de palabras y genera frases coherentes que permiten obtener una clasificación precisa de polaridad de palabras, de la misma forma [21] utiliza una herramienta propia basada en el léxico, SentiText, el cual se enfoca en una párrafo o conjunto de palabras que determinan la intensidad de afecto de la oración, es decir, se calcula el valor global de sentimiento mediante la diferencia entre la suma de negativos y la suma de positivos, multiplicado por la intensidad de afecto.

$$GSV = \frac{(\sum_{i=1}^{5} 2.5*i*N_i + \sum_{i=1}^{5} 2.5*i*P_i)IA}{5*(SL-SN)} \qquad (5)$$

Donde IA representa la intensidad de afecto de la palabra, SL los segmentos léxico y SN los segmentos neutrales.

La negación de palabra mejora de manera notable la calidad de clasificación en cuanto a polaridad de las palabras [19], el uso de un sistema no supervisado permite clasificar de manera correcta estas palabras [16] y [19], para realizar esta clasificación primero se toma la opinión que debe pasar por una división de palabras, caracteres y espacios en blanco (tokenización) luego divide oraciones para luego realizar un análisis morfológico, sintáctico y de dependencias, es aquí donde se realiza el tratamiento de la negación, para esto se utilizan tres recursos: SentiWordNet, iSOl, eSol [16] para calcular la polaridad y clasificar finalmente la opinión.

Se calcula la polaridad de la palabra para que a su vez sea invertida mediante las fórmulas utilizadas en los recursos SentiWordNEt, calcula la suma de la diferencia entre los valores positivos y negativos obtenidos de cada palabra basándose en el sentido más frecuente que se le da.

$$\sum_{i=0}^{n} pos - neg \qquad (6)$$

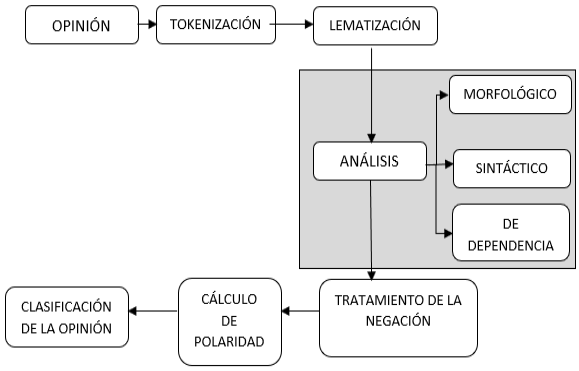


Figura 4.Arquitectura del sistema para tratamiento de negación

iSol y eSol, calculan la suma de los términos, si se clasifican como positivos suma 1 caso contrario resta 1, como resultado eSol arroja los mejores resultados ya que tiene un diccionario más amplio a diferencia de SWN que no realiza una clasificación de manera apropiada; otros autores concuerdan con que eSol devuelve resultados más acertados, estos son [1] ,que utiliza el corpus COAH, cuyo enfoque está basado en las opiniones de los turistas que se hospedan en hoteles, derivando eSol a eSolH (eSol hoteles).

A pesar de que la mayoría de los autores utilizan enfoques basado en Machine Learning y léxicos, [4] utiliza un enfoque tradicional que se basa en combinar tokenizadores para obtener resultados más aproximados a la realidad, el uso de estos es menos complejo que los enfoques léxico y ML indica [4], sin embargo [18] compara el uso de tres herramientas basadas en ML y léxico vs el análisis manual de los textos, en donde indica que cuando se realiza un análisis manual la polaridad tiende a ser negativa y que SAET es la herramienta más efectiva para determinar este tipo de polaridad al ser basada en un enfoque léxico, a su vez Text Analytics y IBMWatson determinan polaridades mayormente positivas y neutras respectivamente.

4. Discusión

SLRP1: *¿Cómo se construyen los corpus o data set que permitan realizar el SA y cuál es el más utilizado?*

De acuerdo con los resultados se determina que el corpus TASS es el más utilizado, este es dividido en secciones según el parea de investigación como política, ciencia, tecnología entre otros, sin embargo, para otras áreas de interés como las hoteleras el corpus COAH es el más apropiado para analizar los sentimientos de los comentarios expresados por los usuarios, en el siguiente gráfico se observa la diferencia de cantidad de palabras de los distintos diccionarios.

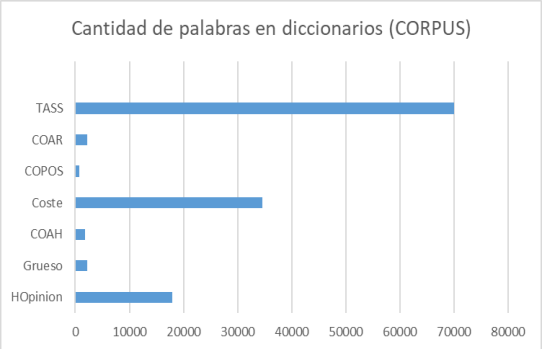


Figura 5. Diccionarios

SLRP2: *¿Cuál es el enfoque más utilizado para SA?*

Aunque el 24% de autores optan por el análisis utilizando enfoque léxico, el 57% han comprobado que utilizar un aprendizaje automático (Machine Learning) devuelve mejores resultados a la hora de clasificar palabras con polaridad positiva, negativa o neutra, el 14% prefiere utilizar un enfoque híbrido y tan solo el 5% utilizan otras técnicas de análisis de sentimientos.

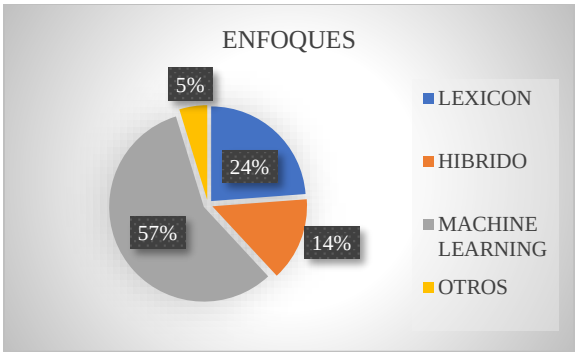


Figura 6. Enfoques

SLRP3: *¿Cuáles son las técnicas que más se utilizan para el SA y bajo que enfoque?*

Las técnicas más utilizadas se encuentran bajo el encuentran bajo el enfoque de Machine Learning, y la técnica que más se utiliza es SVM.

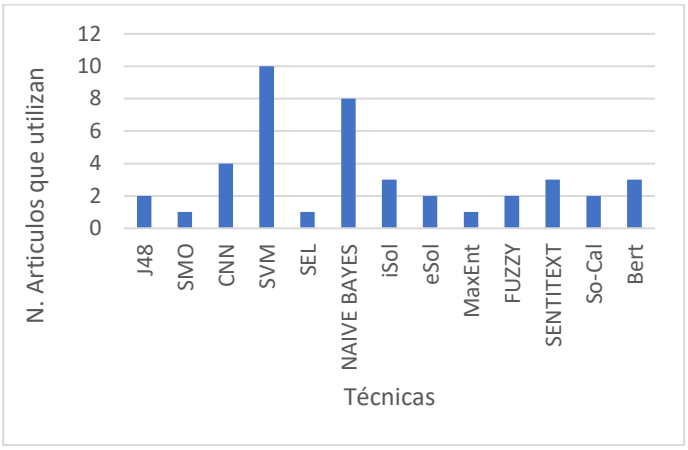


Figura 7. Técnicas

Tabla 9. Técnicas utilizadas por artículos

Técnica	Artículos
J48	[5],[9]
SMO	[9]
CNN	[6].[10].[11].[19]
SVM	[1],[3],[7],[10],[11],[12],[15],[16],[23],[24]
SEL	[24],
NAIVE BAYES	[4],[5],[7],[9],[10],[12],[23],[24]
iSol	[1],[15],[16]
eSol	[1],[16]
MaxEnt	[7]
FUZZY	[2],[18]
SENTITEXT	[12],[14][21]
So-Cal	[3],[23]
Bert	[13],[18],[20]

SLRP4: *¿Qué técnica y enfoque demuestra mayor eficacia?*

De la misma forma se realizaron las comparaciones entre las distintas técnicas de clasificación de palabras, sobresaliendo entre ellas SVM debido a que en cada comparación cuenta una medida de exactitud más alta que otras técnicas, fueron comparadas también otras técnicas con buenos resultados como Naive Bayes, CNN, SentiWord, ANEW,

El enfoque que demuestra mayor eficacia es el Machine Learning ya que se analizan modelos pre-entrenados.

5. Conclusiones

El artículo elaborado fue producto de un análisis de distintas contribuciones realizadas entre el 2009 y 2021, se comparan distintos enfoques y técnicas utilizados para realizar el análisis de sentimientos.

Al inicio de la investigación se encuentra que no existe gran cantidad de información acerca de análisis de sentimientos en español puesto que la mayoría habla sobre el idioma inglés, algunos autores hacen una traducción del idioma para poder realizar el SA, sin embargo, se ha detectado que esto no arroja resultados óptimos en cuanto a la clasificación de palabras.

Actualmente el número de recursos léxicos para realizar análisis de sentimientos para textos en español han incrementado, sin embargo, la mayoría de los autores prefieren utilizar ML ya que en ocasiones este enfoque devuelve resultados más precisos al ser modelos entrenados.

La mayoría de los análisis se han realizado a nivel de documento o frases más que de palabras, ya que de esto permite una mejor percepción del sentimiento que se desea transmitir.

Al final se recomienda un estudio más avanzado de recursos que permitan realizar análisis de sentimientos de textos en español sin recurrir a traducción del idioma, esto permitirá la creación de nuevos modelos que arrojen resultados con mejor precisión.

6. Bibliografía

[1] M. D. Molina-González, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, "Cross-domain sentiment analysis using Spanish opinionated words," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, vol. 8455 LNCS, pp. 214–219. doi: 10.1007/978-3-319-07983-7_28.

[2] P. Rey and D. Castillo, "FUZZ Y SENTIMENT ANALYSIS USING SPANISH TWEETS," 2016. [Online]. Available: https://www.researchgate.net/publication/307598060

[3] J. Brooke, M. Tofiloski, and M. Taboada, "Cross-Linguistic Sentiment Analysis: From English to Spanish," 2009. [Online]. Available: http://garraf.epsevg.upc.es/freeling/

[4] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, and E. A. Villaseñor, "A case study of Spanish text transformations for twitter sentiment analysis," Expert Systems with Applications, vol. 81, pp. 457–471, Sep. 2017, doi: 10.1016/j.eswa.2017.03.071.

[5] M. P. Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, and G. Alor-Hernández, "Sentiment analysis based on psychological and linguistic features for Spanish language," in Intelligent Systems Reference Library, vol. 120, Springer Science and Business Media Deutschland GmbH, 2017, pp. 73–92. doi: 10.1007/978-3-319-51905-0_4.

[6] J. Ochoa-Luna and D. Ari, "Deep neural network approaches for Spanish sentiment analysis of short texts," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11238 LNAI, pp. 430–441. doi: 10.1007/978-3-030-03928-8_35.

[7] M. Del, P. Salas-Zárate, M. A. Paredes-Valverde, J. Limon-Romero, D. Tlapa, and Y. Baez-Lopez, "Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods," 2016. [Online]. Available: http://www.internetworldstats.com/stats7.htm

[8] A. Karpov, R. Potapova, and I. Mporas, Eds., Speech and Computer, vol. 10458. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-66429-3.

[9] J. A. García-Díaz, M. P. Salas-Zárate, M. L. Hernández-Alcaraz, R. Valencia-García, and J. M. Gómez-Berbís, "Machine learning based sentiment analysis on Spanish financial tweets," in Advances in Intelligent Systems and Computing, 2018, vol. 745, pp. 305–311. doi: 10.1007/978-3-319-77703-0_31.

[10] M. A. Paredes-Valverde, R. Colomo-Palacios, M. D. P. Salas-Zárate, and R. Valencia-García, "Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach," Scientific Programming, vol. 2017, 2017, doi: 10.1155/2017/1329281.

[11] I. Segura-Bedmar, A. Quirós, and P. Martínez, "Exploring Convolutional Neural Networks for Sentiment Analysis of Spanish tweets." [Online]. Available: www.sepln.org/workshops/tass/2016/private/evaluate.php

[12] M. T. Martín-Valdivia, E. Martínez-Cámara, J. M. Perea-Ortega, and L. A. Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," Expert Systems with Applications, vol. 40, no. 10, pp. 3934–3942, Aug. 2013, doi: 10.1016/j.eswa.2012.12.084.

[13] J. Pastor-Galindo et al., "Twitter social bots: The 2019 Spanish general election data," Data in Brief, vol. 32, Oct. 2020, doi: 10.1016/j.dib.2020.106047.

[14] D. Salcedo, "Unsupervised Model for Aspect-Based Sentiment Analysis in Spanish." [Online]. Available: https://www.researchgate.net/publication/337228615

[15] O. Araque, I. Corcuera, C. Román, C. A. Iglesias, and J. Fernando Sánchez-Rada, "Aspect based Sentiment Analysis of Spanish Tweets." [Online]. Available: www.en.wikipedia.org/wiki/2014

[16] E. Jiménez Zafra, M. Cámara, M. Valdivia, and M. González, "Negation Scope Identification in Spanish Reviews," pp. 37–44, 2015, [Online]. Available: http://www.redalyc.org/articulo.oa?id=515751523004

[17] M. Navas-Loro and V. Rodríguez-Doncel, "Spanish corpora for sentiment analysis: a survey," Language Resources and Evaluation, vol. 54, no. 2, pp. 303–340, Jun. 2020, doi: 10.1007/s10579-019-09470-8.

[18] I. Utitiaj, P. Morillo, and D. V. Huanga, "Sentiment Analysis Tool for Spanish Tweets in the Ecuadorian Context," Dec. 2020. doi: 10.1145/3446132.3446424.

[19] S. M. J. Zafra, M. Teresa Martin Valdivia, E. M. Camara, and L. Alfonso Urena Lopez, "Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 129–141, Jan. 2019, doi: 10.1109/TAFFC.2017.2693968.

[20] J. Á. González, L. F. Hurtado, and F. Pla, "TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter," Neurocomputing, vol. 426, pp. 58–69, Feb. 2021, doi: 10.1016/j.neucom.2020.09.078.

[21] A. Moreno-Ortiz and C. Pérez Hernández, "Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish," 2013. [Online]. Available: http://tecnolengua.uma.es/sentitext

[22] F. M. Plaza-del-Arco, M. T. Martín-Valdivia, L. A. Ureña-López, and R. Mitkov, "Improved emotion recognition in Spanish social media through incorporation of lexical knowledge," Future Generation Computer Systems, vol. 110, pp. 1000–1008, Sep. 2020, doi: 10.1016/j.future.2019.09.034.

[23] S. M. Jiménez-Zafra, M. T. Martín-Valdivia, M. D. Molina-González, and L. A. Ureña-López, "Relevance of the SFU ReviewSP-NEG corpus annotated with the scope of negation for supervised polarity classification in Spanish," Information Processing and Management, vol. 54, no. 2, pp. 240–251, Mar. 2018, doi: 10.1016/j.ipm.2017.11.007.

[24] O. J. Gambino and H. Calvo, "A comparison between two Spanish sentiment lexicons in the twitter sentiment analysis

task," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 10022 LNAI, pp. 127–138. doi: 10.1007/978-3-319-47955-2_11.
[25] GAMALLO OTERO, P. y s GARCÍA GONZÁLEZ, M. (2000). Técnicas de Procesamiento del Lenguaje Natural en la Recuperación de Información. Centro de Investigación sobre Tecnoloxías da Lingua (CITIUS), 21.
[26] Spanish Billion Word Corpus and Embeddings. (s. f.). Cristian Cardellino. https://crscardellino.ar/SBWCE/

7. ANEXOS

Tabla 10. Matriz de referencia

#Artículo	Nombre	Autor	Año	Idea principal
1	Cross-domain sentiment analysis using Spanish opinionated words	M. D. Molina-González, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López	2014	Utiliza léxicos esol e isol con el corpus COAH, para medir utiliza precisión, recuerdo y F1. Calcula la polaridad con la fórmula: #positivas vs #negativos.
2	Fuzzy sentiment analysis using spanish tweets	P. Rey and D. Castillo	2016	*Utiliza ANEW que es un sistema que se enfoca en el análisis de tweets +Para la construcción del sistema se tuvo en cuenta las fases de extracción, preprocesamiento de textos, identificación del sentimiento y la respectiva clasificación de la opinión utilizando ANEW *Recogen datos del tweet a las 14 hora por 15 minutos mediante una API
3	Cross-Linguistic Sentiment Analysis: From English to Spanish	J. Brooke, M. Tofiloski, and M. Taboada	2009	Para la traducción automática utilizan, dos métodos diferentes. El primero fue un diccionario bilingüe en línea, del sitio www.spanishdict.com. Se extrae la primera definición bajo la categoría sintáctica apropiada, ignorando cualquier caso donde el inglés o el español fueran expresiones de múltiples palabras. El segundo método de traducción automática implica simplemente conectar nuestros diccionarios de inglés en el traductor de Google y analizar los resultados. El tercer método consistía en crear todos los diccionarios desde cero.
4	A case study of Spanish text transformations for twitter sentiment analysis	E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, and E. A. Villaseñor	2017	Utiliza la metodología de análisis exhaustivo, transforma y tokeniza el texto con los corpus INEGI y TASS'15, combina tokenizadores para obtener resultados, indica que no es necesario utilizar otras técnicas puesto que puede ser contraproducente.

5	Sentiment analysis based on psychological and linguistic features for Spanish language	M. P. Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, and G. Alor-Hernández	2017	LIWC (Linguistic Inquiry and Word Count) es una aplicación de software que proporciona una herramienta eficaz para estudiar los componentes emocionales, cognitivos y estructurales contenidos en el lenguaje palabra por palabra
6	Deep neural network approaches for Spanish sentiment analysis of short texts	J. Ochoa-Luna and D. Ari	2018	Explora lacombinación de varias representaciones de palabras (Word2Vec, Glove, Fas-texto y modelos de redes neuronales profundas para clasificar textos breves
7	Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods	M. Del, P. Salas-Zárate, M. A. Paredes-Valverde, J. Limon-Romero, D. Tlapa, and Y. Baez-Lopez	2016	Método de extracción de características híbrido basado en el patrón POS y análisis de dependencia. *El método de clasificación de sentimientos que aquí se propone se divide en cuatro pasos principales: (1)extracción de características, (2) aumentar el espacio de características semánticas, (3) selección de características y(4) entrenamiento del clasificador. *Método EAGLES *Teoría de conjuntos aproximados (RST) *Ganancia de información (IG) mide la relevancia de una característica determinada Entrenamiento de los clasificadores (WEKA)
9	Machine learning based sentiment analysis on Spanish financial tweets	J. A. García-Díaz, M. P. Salas-Zárate, M. L. Hernández-Alcaraz, R. Valencia-García, and J. M. Gómez-Berbís	2018	Se utilizaron tres algoritmos diferentes para procesar los corpus: bow, LIWC y umtextstats. Entonces la combinación de las características obtenidas también se estudió por el enfoque bow y LIWC o umtextstats *La combinación de características umtextstats y bow obtiene el mejor resultado en eltres clasificadores. En concreto, el mejor algoritmo de clasificación fue SMO con una Medida F del 73,2% y la peor fue J48 con una medida F del 62,8%.

10	Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach	M. A. Paredes-Valverde, R. Colomo-Palacios, M. D. P. Salas-Zárate, and R. Valencia-García	2017	El enfoque de clasificación de sentimientos presentado en este trabajo se divide en tres módulos principales: (1) módulo de preprocesamiento, (2) incrustaciones de palabras y (3) modelo CNN ((red neuronal convolucional). * *Utiliza word2vec para incrustación de palabras *Experimenta con tweets positivos y negativos y obtiene una tabla comparativa
11	Exploring Convolutional Neural Networks for Sentiment Analysis of Spanish tweets	I. Segura-Bedmar, A. Quirós, and P. Martínez	2016	*Analiza emoticonos positivos y negativos *Para la tokenización utiliza NLTK (un paquete de Python-2edad para PNL) *Utiliza método similar to Chikersal et al. (2015) para la preclasificación de emoticonos *Se realizó la clasificación de tweets usando scikit-learn Adopta el sistemas TASS * Diferencia entre modelo tweets y cardellino
12	Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches	M. T. Martín-Valdivia, E. Martínez-Cámara, J. M. Perea-Ortega, and L. A. Ureña-López	2013	*Utiliza Machine Learning que es un enfoque supervisado que se basa en usar una colección de datos para entrenar a los clasificadores *Utiliza el corpus muchocine (MC) compuesto por opiniones en español. Consta de 3878 críticas de películas recopiladas del Sitio web de muchocine
13	Twitter social bots: The 2019 Spanish general election data,"	J. Pastor-Galindo et al	2020	Recolecta data set y los clasifica.
14	Unsupervised Model for Aspect-Based Sentiment Analysis in Spanish.	D. Salcedo	2019	• Modelo no supervisado que permite escalabilidad • Sevemal 2016 Utilizan el programa aspectsa desarrollado en JAVA para analizar los sentimientos, el txto pasa por 4 capas para determinar polaridad

15	Aspect based Sentiment Analysis of Spanish Tweets	O. Araque, I. Corcuera, C. Román, C. A. Iglesias, and J. Fernando Sánchez-Rada	2014	Combinan un amplio número de rasgos y léxicos de polaridad para la detección de sentimiento, junto con un algoritmo basado en grafos para la detección de contextos *Reconocimiento de entidad con nombre (NER)módulo *Polaridad de palabras utilizando la negación para comprobar nivel de afectación
16	Negation Scope Identification in Spanish Reviews	E. Jiménez Zafra, M. Cámara, M. Valdivia, and M. González	2015	Realiza tokenización de palabras, realiza análisis morfológico sintáctico y de dependencias, trata la negación con SWN, isol y esol. Calcula polaridad.
17	Spanish corpora for sentiment analysis: a survey	M. Navas-Loro and V. Rodríguez-Doncel	2020	Analiza y describe los distintos corpus que existen.
18	Sentiment Analysis Tool for Spanish Tweets in the Ecuadorian Context	I. Utitiaj, P. Morillo, and D. V. Huanga	2020	Comparan tres técnicas junto al análisis manual, indica que el análisis manual determina mayormente polaridad negativa. Utilizan: Text Anakytics e IBM Watson (ML) SAET (Léxico)
19	Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter	S. M. J. Zafra, M. Teresa Martin Valdivia, E. M. Camara, and L. Alfonso Urena Lopez	2019	Léxico no supervisado para clasificación de polaridad de palabras escrita negativamente.
20	TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter	J. Á. González, L. F. Hurtado, and F. Pla	2021	Bert es capaz de analizar palabras que no expresan una realidad (contextualizadas) TwuilBert supera a BERT multiligüe TwilBert evalua, entrena y ajusta modelos.
21	Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish	A. Moreno-Ortiz and C. Pérez Hernández	2013	Enfoque basado en léxico, utiliza la herramienta Sentitext para analizar el texto,

22	Improved emotion recognition in Spanish social media through incorporation of lexical knowledge	F. M. Plaza-del-Arco, M. T. Martín-Valdivia, L. A. Ureña-López, and R. Mitkov	2020	Utiliza enfoque léxico. Obtiene mejores resultados que Naive Bayes en un 6.15%. Indica que existen recursos léxicos además del inglés que pueden arrojan buenos resultados.
23	Relevance of the SFU ReviewSP-NEG corpus annotated with the scope of negation for supervised polarity classification in Spanish	S. M. Jiménez-Zafra, M. T. Martín-Valdivia, M. D. Molina-González, and L. A. Ureña-López	2018	Corpus sfu
24	A comparison between two Spanish sentiment lexicons in the twitter sentiment analysis task	O. J. Gambino and H. Calvo	2016	*Utiliza la plataforma TASS, y analiza el corpus con 2 léxicos españoles distintos, basados en reglas y supervisados, se obtiene un mejor resultado con el léxico LIWC que con SEL, se crea una matriz de confusión en la cual se muestran los resultados obtenidos de las clases neutrales, positivas y negativas, en donde se comprueba que si el entramiento e intersección del vocabulario pasa el 50% éste genera un mejor impacto en el corpus y por lo mismo un mejor resultado.