UNIVERSIDAD CATÓLICA DE LA SANTÍSIMA CONCEPCIÓN
Facultad de Ingeniería
Ingeniería Civil Informática

ANÁLISIS DE POLARIDAD EN TWITTER, UTILIZANDO RASGOS DE SUPERFICIE, SEMÁNTICOS Y LEXICONES

Iván Leonardo Castro Montero

INFORME DE PROYECTO DE TÍTULO PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

Profesor Guía
Jose Ignacio Abreu Salas

Concepción, Julio 2017

Resumen del proyecto de título

En esta investigación se propone un enfoque que combina rasgos de superficie, lexicones y rasgos semánticos. Para esto se realizó un sistema el cual genera 160 modelos distintos, que demuestran que este enfoque logra alcanzar resultados que permiten ser competitivos. Se investiga además la efectividad de ciertos atributos considerados como dinámicos porque su número depende del *Corpus*, tales como *Chargrams*, *Ngrams* y *Clusters*, los que para este estudio en algunos casos resultan ser prescindibles.

Se postula además la utilización de sentidos, como rasgos semánticos y el estudio de la polaridad para rasgos de superficie.

Abstract

This research proposes an approach that combines features of surface, lexicons and semantic features. This was a system which generates 160 different models, which show that this approach can achieve results that allow you to be competitive. In addition the effectiveness of certain attributes considered as dynamic because their number depends on the Corpus, such as Chargrams, Ngrams and Clusters, that happen to be dispensable for this study in some cases investigated. It is postulated in addition the use of senses, as semantic features and the study of the features of surface polarity.

Dedicatoria

Terminar con una etapa con tanto significado en mi vida, me llena de emociones. Saber que el esfuerzo tuvo su recompensa, me hace pensar que por muy lejos se vea la meta no hay que bajar los brazos. Los buenos momentos sirven para seguir de pie, los malos, para hacer más fuerte los sueños y objetivos.

Quiero agradecer a mis padres Ruperto Castro y Gabriela Montero por acompañarme en todo momento, inculcarme valores y principios que me forjaron y forjan como persona, estudiante y profesional.

Olvidarme de mis abuelos, familia, polola, amigos, compañeros y de todos los que forman y formaron parte de mi vida es algo que no me puedo permitir, como también dejar de lado a quienes me enseñaron sus conocimientos desde el momento que pisé mi primera aula de clases; en fin, agradecerle a la vida y a Dios por ponerlos en mi camino.

Iván Castro Montero.

Índice

Índice de figuras

Índice de tablas

Capítulo 1

1. Introducción

1.1. Presentación del tema

En la actualidad las redes sociales pasan a formar parte de nuestro diario vivir, convirtiéndose en una herramienta eficaz para masificar información respecto a opiniones, sentimientos o emociones en temas personales o que afecten a nivel de sociedad.

Twitter es una forma expedita para compartir información dadas sus características, tales como, etiquetado *hashtag*, delimitación de caracteres, uso de emoticones, entre otras, permitiendo rápida propagación de información, que se traduce en la masificación de temas, siendo esto una gran oportunidad para el análisis de sentimientos debido al amplio abanico que abarca.

Existe una diversidad de enfoques para asignar el sentimiento que prevalece dentro de un *tweet*, considerando una clasificación de éste como positivo, negativo o neutro. En el análisis propuesto, se buscó dar solución al problema mediante un enfoque que utiliza rasgos de superficie (rasgos léxicos, sintácticos y morfológicos), lexicones y agregando al estudio de la polaridad rasgos semánticos (Ngrams de sentidos).

El análisis de sentimientos, al ser parte del procesamiento del lenguaje natural (PLN), cobra gran importancia debido a que permite delimitar la polaridad de la información de forma automática. Esto significa que se puede saber si el tipo de información compartida en las redes sociales contiene o no algún tipo de sentimiento.

En este estudio, para la predicción de la polaridad, se consideró la clasificación taxonómica de sentimientos como positivo, negativo o neutro, donde este último hace referencia a la ausencia de sentimiento. Conocer esta característica de los tweets permite

generar distintos estudios, lo que despierta el interés de una serie de actores sociales, ya que a través de este medio se puede obtener tendencias, posibles comportamientos o niveles de aceptación respecto a una marca comercial o situaciones en específico.

1.2. Justificación del problema

El estudio se abocó en proponer un enfoque que permita dar polaridad a tweets de forma automática, lo que abre un gran abanico de posibilidades para realizar aplicaciones en diversos ámbitos tanto políticos, educacionales, económicos, sociales, entre otros.

Es importante señalar que no existe una solución concreta, por tanto, se pueden generar nuevos modelos que asignen polaridad.

1.3. Delimitación del Problema.

Se analizarán Corpus de tweets para el entrenamiento y predicción de polaridad, facilitados por el Workshop internacional SemEval, del período comprendido entre los años 2013 al 2017, en el idioma inglés.

1.4. Objetivo general.

Proponer un enfoque que combine rasgos de superficie, lexicones y rasgos semánticos, para determinar la polaridad de un tweet.

1.5. Objetivos específicos.

1. Revisar bibliografía sobre los diferentes enfoques para el análisis de polaridad.

2. Definir un sistema para asignar polaridad a tweets, que integre rasgos de superficie, lexicones y rasgos semánticos.

3. Validar experimentalmente la propuesta.

1.6. Metodología

1. Mediante revisión bibliográfica, en las principales bases de datos científicas, se recopilará información referente al tema a investigar, procurando obtener conceptos claves y además generar el estado del arte.

2. Se investigará sobre rasgos que aportarán información semántica, enfoques basados en rasgos de superficie y lexicones para la asignación de polaridad, realizando un estudio comparativo.

3. Se realizarán experimentos comparativos sobre los Corpus facilitados por SemEval en los distintos años de competencia, con la finalidad de analizar sus resultados.

Capítulo 2

2. Marco teórico

2.1. Polaridad

Se define la polaridad como la medida utilizada para clasificar el sentimiento que prevalece dentro de un tweet. La polaridad puede ser positiva, negativa o neutra, dependiendo del sentimiento imperante en un tweet. Por ejemplo, cuando un tweet presenta ausencia de sentimientos, se clasifica como polaridad neutra; por el contrario, si éste tiene el sentimiento positivo como predominante, se considerará con polaridad positiva, lo mismo ocurre cuando predomina el sentimiento negativo, se clasificará con polaridad negativa (HLTCOE, 2013).

2.2. Corpus

Corpus o en su plural corpora, es una colección de textos escritos o transcripciones del lenguaje oral para cierto idioma, que en la actualidad tienen al menos un millón de palabras y, por lo general, se almacena en un formato electrónico legible por máquinas.

Un Corpus puede ser un banco de pruebas para las hipótesis y se puede utilizar para añadir una dimensión cuantitativa en muchos estudios lingüísticos. La mayoría de los corpora contienen información adicional a los textos que las componen, como la información sobre los propios textos, etiquetas POS para cada palabra, entre otros (Hunston, 2006).

2.3. Stopwords

Las *stopwords*, también llamadas palabras vacías, representan el ruido en el proceso de recuperación de los términos más usados; generalmente suelen ser conectores, preposiciones y artículos. Además, su detección y posterior eliminación permite reducir el tamaño de almacenamiento de la colección indexada (Indurkhya and Damerau, 2010).

2.4. Token

Son elementos del texto que se identifican en el procesamiento del Corpus. Son cadenas de caracteres que se entienden como unidades indivisibles. Pueden mezclar caracteres y símbolos como números, abreviaturas, acrónimos, fechas, entre otros (Indurkhya and Damerau, 2010).

2.5. Ngrams

Se llaman Ngrams a una subsecuencia de n palabras consecutivas en una secuencia dada. Podemos identificar frecuencias de palabras; si hablamos de una palabra estaríamos frente a un Unigrams, si el caso fuese una frecuencia de pares de palabras adyacentes, se denominaría Bigrams y así sucesivamente (Nugues, 2006).

2.6. Etiquetas POS

Las etiquetas POS o etiquetas gramaticales consideran las funciones sintácticas y morfológicas de las palabras dentro de la oración. Para el procesamiento del lenguaje natural el etiquetado POS es la anotación automática de palabras con categorías gramaticales (Nugues, 2006).

2.7. Lexicón

Es un conjunto de palabras, similar a un diccionario, que a menudo cubren un dominio en particular, algunos se centran en todo un lenguaje, como inglés, francés o alemán, mientras que otros se especializan en un área en particular (Nugues, 2006).

Para el análisis de sentimientos en Twitter existen lexicones basados en características propias de la red social, por lo que se pueden encontrar lexicones basados en emoticones, hashtag, etcétera. Estos tendrán asociado su polaridad dependiendo de la taxonomía que se utilizó para clasificar su sentimiento.

2.8. Synset

Synset es la unidad básica con la que se estructura la base de palabras en inglés WordNet. Este concepto engloba sustantivos, verbos, adjetivos y adverbios, los cuales son agrupados en un conjunto de sinónimos cognitivos, conectados por relaciones conceptuales semánticas y léxicas (Fellbaum, 2005).

2.9. SentiWordNet

Este recurso léxico utiliza los Synset de la base de palabras en inglés WordNet. SentiWordNet permite identificar la polaridad de los Synset de forma automática dependiendo de los grados de positividad, negatividad y neutralidad que estos posean (Baccianella et al., 2010).

2.10. Aprendizaje automático

Se dice que un programa de computadora es capaz de aprender a partir de cierta experiencia (E) respecto a una determinada tarea (T) y una medida de desempeño

(P) si su desempeño en la tarea T se incrementa con la experiencia E(Mitchell, 1997). Existen tres algoritmos que presentan buen desempeño en la categorización del texto: Naive Bayes(NB), Máxima Entropía(MaxEn) y SVM(Support Vector Machine) (Pang et al., 2002), siendo este último aplicado para el enfoque propuesto. El SVM permite clasificar datos lineales y no lineales mediante mapeos a los datos de entrada, otorgando características a una dimensión mayor, encontrando un hiperplano para una separación, formando frontera de decisión. La utilidad radica en la posibilidad de separar datos en intervalos de clases para determinar a cuál pertenece, dando respuesta a complejos límites de decisión (Han and Kamber, 2011).

2.11. Algoritmo de Lesk

Según este algoritmo, se puede desambiguar el sentido de una palabra en un contexto, identificando el total de sentidos de palabras relacionadas en el texto (Lesk, 1986). El único recurso requerido por el algoritmo es un conjunto de entrada de un diccionario.

2.12. CMU Pos-Tagging Tool

Esta herramienta proporciona mil cluster o agrupaciones de palabras, generados por el algoritmo Brown Cluster que analizó 56 millones de tweets (Mohammad et al., 2013). Además, permite tokenizar el tweet, asignando a cada *token* su etiqueta POS.

Como es una herramienta específica para twitter, tiene la capacidad de reconocer si los tokens son emoticones, hashtag, URL y otras características propias de la red social twitter(Gimpel et al., 2011).

2.13. Test de rangos signados de Wilcoxon

Es una prueba no paramétrica que es usada para hacer pruebas de hipótesis acerca de la mediana, comparando dos muestras relacionadas. No necesita de una distribución específica; lo único que se requiere es que la variable sea continua y sean observaciones pareadas.(Gómez-Gómez et al., 2003).

Capítulo 3

3. Estado del arte

Para conocer las tendencias y estrategias de predicción de polaridad sobre Corpus, se consultaron diversas fuentes de información, rescatando así modelos y propuestas.

3.1. Estudios de polaridad y técnicas

El análisis de sentimiento, específicamente la predicción de polaridad, puede ser aplicada dentro de cualquier documento de texto electrónico u oraciones en particular.

En un comienzo la tendencia de los trabajos de investigación, que estudian el análisis de sentimiento, lo hacen utilizando una taxonomía que sólo clasifica al sentimiento como positivo o negativo. Existen dos estudios que se destacan siguiendo esta premisa y que además plantean dos formas de abordar esta problemática: El primero analiza críticas de películas generadas en la Web, específicamente de la base de datos de IMDb, utilizando técnicas de aprendizaje automático, basadas en tres algoritmos, estos son, Naive Bayes (NB), Máxima Entropía (MaxEn) y Support Vector Machine(SVM). Los anteriores algoritmos mostraron un desempeño similar obteniendo resultados con pequeñas diferencias, donde SVM estuvo por sobre de los demás. La extracción de atributos o características en la etapa de aprendizaje de éstos, se basó en la utilización de unigrams y bigrams para su análisis. Otro punto a tomar en consideración de este estudio, es la problemática generada por opiniones que presentaban una ausencia de sentimiento, dejando abierto el debate si la complejidad de clasificar las opiniones con la taxonomía propuesta era suficiente para llegar a un buen resultado (Pang et al., 2002).

El segundo estudio también analiza críticas generadas en la web; esta vez el dominio a estudiar fue automóviles, películas, bancos y destinos vacacionales. En este estudio se pueden distinguir tres etapas: la primera se enfoca en extraer frases que contienen adjetivos o adverbios, luego se realiza una estimación de la orientación semántica de las frases, la que es calculada por el algoritmo Pointwise Mutual Information -Information Retrieval (PMI-IR) y, por último, se clasifica la opinión basándose en el promedio de la orientación semántica de las frases (Turney, 2002).

Si bien en los estudios mencionados, no se utilizó como opción el uso de una taxonomía que considera la premisa que algunas frases presentan ausencia de sentimiento, en 2006 (Koppel and Schler, 2006), muestra la importancia de considerar un sentimiento como neutro en representación de la ausencia de sentimientos, y cómo ésta mejora la predicción y ayuda a los algoritmos en el proceso de aprendizaje.

3.2. Nacimiento de Twitter y posterior foco de estudio

Desde la creación de Twitter (21 de marzo del 2006) a la actualidad, múltiples estudios han sido realizados debido a su alto impacto. Inspirados por dicha importancia, empresas y organizaciones buscan la forma de obtener la información que ahí se comparte para usarla con diversos fines.

En los años 2009-2010 se empezaron a generar gran cantidad de artículos relacionados con la problemática de predicción de polaridad en Twitter, enfocado en distintos ámbitos, como por ejemplo la investigación de (Jansen et al., 2009) cuyo objetivo tiene relación con la utilización de Twitter como un "boca a boca" electrónico para el intercambio de opiniones de los consumidores acerca de marcas y servicios determinados. Para este trabajo se utilizaron dos tipos de clasificadores, uno automático y otro manual, con la finalidad de comparar la precisión del primero. Para el caso del clasificador

automático que sigue un modelo de NB, se recogieron datos de tweets con la herramienta SUMMIZE para todas las marcas en un período de 13 semanas. Luego, se calculó la clasificación de cada tweet. Esta herramienta utiliza un lexicón de aproximadamente 200 mil unigrams y bigrams de palabras y frases que tienen una distribución de probabilidad para determinar el sentimiento de la marca para un determinado período. El clasificador de SUMMIZE se capacitó con casi 15 millones de puntos de vista sobre temas que van desde las películas a la electrónica, con el fin de determinar cómo las personas utilizan adjetivos para manifestar sus opiniones de forma positiva o negativa en línea. Este estudio se midió con la escala de Likert (clasificación: *wretched, bad, so-so, swell, and great*).

Twitter pasa a constituir así una herramienta en línea, tanto para los clientes como para las corporaciones, debido al fuerte impacto que produce en las estrategias de mercados. Otro ámbito donde se utilizó Twitter fue para el estudio de la predicción de elecciones, realizado por (Tumasjan et al., 2010), donde se analizaron 104.003 tweets donde mencionaban o hacían referencia a un partido político o un político en las semanas previas (específicamente entre el 13 de agosto al 19 de septiembre del 2009) a la elección federal del parlamento nacional en Alemania a realizarse el día 27 de septiembre del 2009.

El proceso de generación de tweets fue aumentando mientras más se acercaba la fecha de las elecciones, estos tweets fueron analizados por el software LIWC, una herramienta que calcula el grado en que las personas utilizan diferentes categorías de palabras relacionadas con los procesos psicolingüísticos, para lo cual los tweets tuvieron que ser traducidos del alemán al inglés, para así poder ser procesados por el lexicón interno del software. Este estudio presentó varias limitaciones, ya sea idiomáticas como también la no utilización de características de Twitter como hashtag y emoticones; a pesar de eso, se llegó a la conclusión que la herramienta se puede utilizar como

complemento para conocer opiniones y predecir resultados.

3.3. SemEval y su tarea de análisis de sentimientos en Twitter

En 2013 en el Workshop SemEval se postuló la tarea de análisis de sentimientos en Twitter, la que tuvo gran aceptación por la comunidad científica y hasta el día de hoy se sigue realizando. En ésta se generaron varias subtareas, dentro de las cuales centraremos nuestro punto de atención en "Message Polarity Classification", tarea en la cual se debe predecir un corpus de tweets sin etiquetar, con una taxonomía para la polaridad que se mide en una escala positivo, negativo y neutro. Sin embargo, el valor de la predicción de polaridad neutra no se considera para calcular el F_{score}. Los valores positivos y negativos se calculan con la métrica Precision-Recall, por ejemplo Precision (P). Para los tweets positivos, se calcula buscando el número de tweets que el sistema predijo de forma correcta y se divide por el número total de tweets que se predijo como positivo. Para el Recall (R), en los positivos, se calcula el número de tweets que el sistema predijo correctamente y los divide por el total de tweets positivos del Corpus. Teniendo estos datos se calcula un F_{pos} como se muestra en la ecuación 1. Para el caso de los tweets negativos, se realiza el mismo procedimiento. Finalmente se calcula un promedio entre ambos como se muestra en la ecuación 2.

$$F_{pos} = 2\frac{P_{pos}R_{pos}}{P_{pos} + R_{pos}} \tag{1}$$

$$F_{Score} = \frac{F_{pos} + F_{neg}}{2} \tag{2}$$

3.3.1. SemEval 2013

La tarea tuvo una convocatoria de 38 equipos dentro de los cuales destaca, NRC-CANADA (Mohammad et al., 2013), quienes entrenaron un algoritmo de aprendizaje

automático (SVM), con la extracción de una gran cantidad de características y el uso de lexicones, obteniendo el primer lugar con un 69,02 %. El modelo consiste en preprocesar los tweets, donde reemplazan las URL, usuarios y menciones, para que no genere ruido en el proceso de extracción de características. Una vez preprocesados los tweets, se tokenizan y se les asignan etiquetas POS, luego se extraen atributos como por ejemplo, Ngrams,Ngrams de caracteres(Chargrams), cantidad de palabras en mayúscula, cantidad de hashtag,etcétera. Los Ngrams, en este caso los Unigrams y Bigrams, son analizados por lexicones de sentimientos, los cuales contienen polaridades. Los tweets ahora se representan como vectores de características y se procede a entrenar el SVM.

Otros equipos como GU-MLT-LT (Wijksgatan and Furrer, 2013) y KLUE (Proisl et al., 2013), utilizan un enfoque similar, donde se preprocesan los tweets, se extraen características, que en estos casos es menor a la cantidad considerada por NRC-CANADA, se utilizan lexicones, pero se entrenan algoritmos diferentes, GU-MLT-LT entrena un algoritmo de Gradiente Estocástico , obteniendo el segundo lugar con un 65,27 %. Por otro lado KLUE entrenó un clasificador de MaxEn, obteniendo el quinto lugar con un 63,03 %.

También existen equipos que prescinden del uso de algoritmos de aprendizaje automáticos para la predicción, como es el caso de TERAGRAM (Reckman et al., 2013) y SSA-UO (Ortega Bueno et al., 2013), quienes basaron su modelo en reglas.TERAGRAM genera distintas taxonomías para evaluar el tweet y dependiendo de el peso que tengan éstas, se predice su polaridad. Con este enfoque obtuvieron el tercer lugar con un 64,86 %. En el caso de SSA-UO, además se considera el proceso de desambiguar las palabras para conocer el contexto de éstas,obteniendo un 50,17 % que lo deja fuera del top 20 en el ranking.

Otro enfoque que estuvo presente en esta versión fue la del equipo UOM (Negi and Rosner, 2013), quienes entrenaron un algoritmo de aprendizaje automático (NB)

y utilizaron ESA (Explicit Semantic Analysis), con la premisa de que si un texto con polaridad conocida es similar a nivel semántico a otro sin conocer, existe gran probabilidad que compartan la misma polaridad; la decisión final se toma consultando los resultados obtenidos por NB y ESA.

3.3.2. SemEval 2014

La tarea tuvo una convocatoria de 44 equipos dentro de los cuales destaca TEAMX (Miura et al., 2014),el modelo presentado por este equipo se basó en el expuesto por NRC-CANADA el año 2013, pero con una menor cantidad de atributos y más uso de lexicones, además de entrenar un algoritmo de Regresión Logística , en vez de un SVM. Al preprocesamiento agregaron una instancia para la corrección ortográfica y utilizaron dos etiquetadores POS; uno para textos formales (Stanford POS) y otro especializado en Twitter(CMU Pos-Taggin Tool). Además, tomaron en consideración la importancia del sentido de las palabras, ya que las desambiguan antes de consultar SentiWordNet. El equipo obtuvo el primer lugar con un 70,96 %.

NRC-CANADA(Zhu et al., 2014), volvió a presentar el mismo modelo de predicción, pero con modificaciones a su enfoque para la detección de negación y mejoras en los lexicones, obteniendo un cuarto lugar con un 69,85 %.

Otro enfoque que se dio a conocer este año, fueron los que utilizan Deep Learning para la predicción de polaridad. Los equipos COOOLL (Tang et al., 2014) y THINK POSITIVE (dos Santos, 2014), extraen representationes vectoriales de las palabras llamadas *embeddings* y utilizan redes neuronales para la predicción de polaridad, obteniendo el segundo lugar con un 70,14 % y un décimo lugar con un 67,14 %, respectivamente. Para el caso del equipo Coooll su enfoque además replicó y utilizó las características de NRC-CANADA para enriquecer su modelo.

Siguiendo la línea de enfoques que utilizan como parte de su modelo la similitud semántica, el equipo ukpdipf (Flekova et al., 2014) genera su modelo en base a ESA acompañado de un Sequential minimal optimization (SVM-SMO), obteniendo un vigésimo primer lugar con un 63,77 %. El equipo sail (Malandrakis et al., 2014), mezcla un enfoque que utiliza la extracción de características y el uso de lexicones, con un modelo que mezcla métricas de similitud semánticas, modelo el cual participo el año 2012 en la tarea de similitud semántica en SemEval, este equipo obtuvo un séptimo lugar con un 67,04 %.

3.3.3. SemEval 2015

La tarea tuvo una convocatoria de 40 equipos dentro de los cuales destaca WEBIS (Büchner and Stein, 2015), investigación que consistió en replicar cuatro modelos utilizados en años anteriores, que siguen el mismo enfoque, preprocesar los tweets, extraer características, uso de lexicones y el entrenamiento de un algoritmo. Estos equipos fueron NRC-CANADA 2013, GU-MLT-LT 2013, KLUE 2013 y TEAMX 2014.

El modelo de predicción final utiliza las predicciones entregadas por los modelos antes mencionados y a través de un promedio, se realiza la predicción final, obteniendo así el primer lugar con un 64,84 %.

Los enfoques con la utilización de Deep Learning y *embeddings* se hacen más comunes este año destacando el funcionamiento de UNITN (Severyn and Moschitti, 2015), quienes obtuvieron el segundo lugar con un 64,59 %. La contribución de este trabajo se basa en los pesos de inicialización de la red neuronal, que luego se entrena con los datos disponibles para este año, para la predicción de polaridad. Otros equipos que siguieron el mismo enfoque fueron ECNU (Zhang et al., 2015) y CIS-POSITIVE (Ebert et al., 2015) , aunque estos modelos mezclan la utilización de redes neuronales con

la extracción de características y uso de lexicones, entrenando un SVM, obteniendo así el décimo octavo y décimo noveno lugar con un 59,72 % y un 59,57 % respectivamente.

GTI (Fernández-Gavilanes et al., 2015), muestra un enfoque basado en reglas con el uso de lexicones, preocupándose de la sintaxis de los tweets, sus intensificadores, disminuidores y modelos de negación. Este estudio obtuvo un vigésimo primer lugar con un 58,95 %.

3.3.4. SemEval 2016

La tarea tuvo una convocatoria de 34 equipos dentro de los cuales destaca SWISS-CHEESE (Deriu et al., 2016). El modelo se basa en dos capas de redes neuronales donde cada una de estas se diferencian por el tipo de filtros utilizados para la extracción de *embeddings* dentro de un Word2vec. Luego esta representación de vectores resultantes, se analiza en un algoritmo Random Forest para predecir, obteniendo el primer lugar con un 63,3 %.

Los enfoques basados en Deep Learning, con la utilización de Redes Neuronales, Redes Convolucionales(CNN) y la extracción de word embedding, son los que mejor resultado obtuvieron , ejemplo de esto son los equipos SENSEI-LIF (Rouvier and Favre, 2016), UNIMELB (Xu et al., 2016),INESC-ID (Amir et al., 2016),INSIGHT-1 (Ruder et al., 2016), quienes obtuvieron segundo, tercer, cuarto y octavo lugar, con 63 %, 61,7 %, 61 % y 59,3 % respectivamente.

Capítulo 4

4. Descripción del sistema

Para conocer el desempeño del enfoque propuesto, se generó un sistema con distintos modelos de atributos que cumpliesen con sus características.

4.1. Selección de modelos

Se buscó proponer un enfoque que tuviese como características el uso de rasgos de superficie y lexicones para enriquecerlo con rasgos semánticos. Para cumplir con esta premisa se estableció un modelo base que incluye los rasgos de superficie y lexicones.

El modelo que es más referenciado en la revisión bibliográfica utilizando dichas características es el de NRC-CANADA, equipo que para muchos autores marca el estado del arte, por su selección de atributos, lexicones y algoritmo de aprendizaje automático.

El modelo NRC-CANADA (Mohammad et al., 2013) considera son los siguientes atributos:

1. Ngrams(Ngrams de palabra): presencia o ausencia de secuencias continuas de uno, 2 , 3 y 4 tokens.

2. Chargrams(Ngrams de caracter): presencia o ausencia de secuencias continuas de 3, 4 y 5 caracteres.

3. All-caps(Palabras en mayúscula): número de palabras donde todos sus caracteres se encuentran en mayúscula.

4. POS: número de ocurrencia de etiquetas POS.

5. Hashtag: número de hashtag presentes dentro del tweet.

6. Lexicones:En total son cinco lexicones de polaridad, tres de estos creados manualmente, NRC Emotion Lexicón con 14.000 palabras (Mohammad and Turney, 2013),the MPQA Lexicón con 8.000 palabras (Wilson et al., 2005) y the Bing Liu Lexicón con 6.000 palabras (Hu and Liu, 2004). Los otros dos creados de forma automática, NRC hashtag Sentiment Lexicón con 54.129 unigrams,316.531 bigrams y 308.808 pares de palabras no contiguas y NRC Sentiment140 Lexicón con 62.468 unigrams 677.698 bigrams y 480.010 pares de palabras no contiguas.

   Por cada token ω y emoción o polaridad ρ, se usa el $score(\omega, \rho)$ para determinar:

   - cantidad de tokens en el tweet con $score(\omega, \rho)>0$.
   - total score $=\sum_{\omega \in tweet} score(\omega, \rho)$.
   - score máximo $= max_{\omega \in tweet} score(\omega, \rho)$.
   - score del último token en el tweet con $score(\omega, \rho)>0$.

7. Punctuation(Puntuación):

   - el número de secuencias contiguas de signos de exclamación, signos de interrogación y ambos signos de exclamación e interrogación.
   - si el último token contiene signo de exclamación o interrogación.

8. Emoticons(Emoticones):

   - presencia o ausencia de emoticones positivos o negativos en cualquier posición del tweet.
   - si el último token es un Emoticon positivo o negativo.

9. Elongated words(Palabras alargadas): número de palabras donde se repite una letra más de dos veces.

10. Cluster:

    - presencia o ausencia de tokens dentro de los 1000 clusters, proporcionados por CMU Pos-Taggin Tool.

11. Negation(Negación): número de contextos negados, se define contexto negado a un segmento de tweet que comienza con una palabra de negación y termina con uno de los signos: ‘,’, ‘.’, ‘:’, ‘;’, ‘!’, ‘?’, por ejemplo,shouldn’t. Un contexto negado afecta a las características del uso de lexicón.Se le agrega el sufijo ’_NEG’ a cada palabra después de la negación, por ejemplo, la palabra “perfect” se convierte en “perfect_NEG”. Para la generación de Ngrams no se considero la negación a diferencia de WEBIS.

El sistema de NRC-CANADA se puede separar en cuatro partes:

1. Preprocesamiento de los tweet.
2. Extracción de rasgos.
3. Etapa de aprendizaje o entrenamiento del SVM.
4. Predicción.

Como el sistema no está disponible por los autores originales, varios equipos de investigación han replicado su modelo basándose en los artículos que hablan sobre éste. En el año 2015 el equipo WEBIS replicó el trabajo de NRC-CANADA 2013 y de 3 equipos más, dejándolo en una biblioteca pública de uso libre.

Analizando el sistema de WEBIS, se separó el sistema de NRC-CANADA y se utilizó como base.

Las cuatro etapas mencionadas anteriormente, se separaron en dos procesos:

1. Entrenamiento: Esto incluye el preprocesamiento de los tweets, la extracción de atributos y la etapa de entrenamiento del SVM.

2. Test: Se predice la polaridad de los tweets, con el modelo ya entrenado.

Esta separación en la librería se realiza por la flexibilidad que se obtiene al momento de experimentar, esto quiere decir que se puede entrenar un modelo y realizar distintas predicciones sobre ese modelo, sin la necesidad de ejecutar la librería entera. Luego se experimentó con los Corpus de train y test disponibles de la competencia SemEval 2013.

Al analizar los rasgos que se extraen de esta librería, cae en evidencia, la pérdida de información por no analizar los tweet en UTF-8, algunas palabras y emoticones, no son considerados. Por otro lado, la captura de algunas expresiones regulares, generan ruido por no cumplir con su función en totalidad.

4.2. Modificaciones al sistema de NRC-WEBIS

Se reemplazó la expresión regular que captura emoticones (Figura1), por la condición que utiliza la etiqueta de CMU Pos-Taggin Tool para identificar emoticones (Figura2). Dicho cambio fue considerado debido a que no se capturaban algunos emoticones, por el hecho de que ésta, fue elaborada para la captura de emoticones que consideraron como positivos o negativos y de uso frecuente(Potts, 2011), acortando así, la amplia gama de emoticones existentes. En esta investigación todos los emoticones representan algún tipo de información y la expresión regular utilizada por CMU Pos-Taggin Tool, no se limita a una clasificación dual.

```
protected Set<String> getEmoticons(Tweet tweet){
    Set<String> emoticons = new HashSet<String>();
    String emoticon_string =
            "(?:[<>]?[:;=8][\\-o\\*\\']?[\\)\\]\\(\\[dDpP/\\:\\}\\{@\\|\\\\]"
            + "|[\\)\\]\\(\\[dDpP/\\:\\}\\{@\\|\\\\] "
            + "[\\-o\\*\\']?[:;=8][<>]?)";
    Matcher m = Pattern.compile(emoticon_string).matcher(tweet.getTweetString());
    while (m.find()){
            emoticons.add(tweet.getTweetString().substring(m.start(), m.end()));
            if(m.end() == tweet.getTweetString().length()) tweet.setLastEmoticon(true);
    }
    tweet.setEmoticons(emoticons);
    return emoticons;
}
```

Figura 1: Expresión regular inicial, propuesta por Christopher Potts (Potts, 2011).

```
protected Set<String> getEmoticons(Tweet tweet){
    Set<String> emoticons = new HashSet<String>();
    for(TaggedToken tt:tweet.getWordList()){

        if(tt.tag.equals("E")){
            tt.token = contraerPalabra(tt.token,true);
            emoticons.add(tt.token);
        }
    }
```

Figura 2: Condición de captura de emoticones.

Como CMU Pos-Taggin Tool analiza los textos, en este caso los tweets en UTF-8, se generó un flujo donde los tweets se guardan en dicho formato, lo que evitó la perdida de información considerada como relevante.

Para la captura de palabras alargadas se cambió la expresión regular (Figura3),por una que sólo acepta caracteres del abecedario (Figura4) y se complementa con la función ContraerPalabra (Figura5). Su labor es normalizar caracteres consecutivos repetidos más de 2 veces ; por lo tanto, la expresión regular actúa como filtro y la función detecta si la palabra es alargada. Esto se logró comparando la palabra inicial con la procesada por la función; si estas son distintas, se considera como palabra alargada. Por ejemplo,

la palabra “cooool” normalizada queda como “col” y, al momento de ser comparada, se considera como una palabra alargada, evitando generar ruido para el SVM con la captura de números o símbolos indeseados.

```
    private int getElongatedCount(String tweetString){
    int elongatedWords = 0;
    for (String word: tweetString.split("[\\p{P} \\t\\n\\r]")){
            Matcher m = Pattern.compile("(.)\\1{2,}").matcher(word);
            if(m.find()) elongatedWords++;
    }
    return elongatedWords;
}
```

Figura 3: Expresión regular palabras alargada inicial.

```
private int getElongatedCount(Tweet tweet){
        tweet.setPalabras_alargadas(new HashSet<>());
        int elongatedWords = 0;
        String word;
        String tweetString = tweet.getTweetString();
        for (String w: tweetString.replaceAll("[^a-zA-Z]", " ")
                .replaceAll("(\\ ){2,}", " ").trim().split(" ")){
                word = contraerPalabra(w,true);
                //if (!word.equals(tt.token)){
                if (!word.equals(w)){
                        elongatedWords++;
                }
        }
        return elongatedWords;
    }
```

Figura 4: Expresión regular palabras alargadas modificada.

La normalización de palabras alargadas puede presentar problemas, debido a que en el idioma inglés, existen palabras donde un caracter se repite dos veces como es el caso de “food”. Para ello se optó por la opción de dejar la palabra con el caracter repetido una ó dos veces. Para conocer cuál es la mejor opción de normalizado se consulta por la palabra con el caracter repetido dos veces, en un listado de 354.985 palabras en

inglés[1]. Si ésta no se encuentra, se vuelve a normalizar dejando el caracter repetido en uno,consultando nuevamente por la palabra; si ésta no se encuentra de ninguna de las dos formas, se guarda con la última opción de normalizado.

```
protected String contraerPalabra(String n, boolean t){
Matcher m = Pattern.compile("(.)\\1{2,}").matcher(n);
String tempString = n;
while(m.find()){
    if(t){
   n = tempString.substring(0, m.start()+1) + tempString.substring(m.end());
    m = Pattern.compile("(.)\\1{2,}").matcher(n);
   tempString = n;
}else{
    n = tempString.substring(0, m.start()+2) + tempString.substring(m.end());
    m = Pattern.compile("(.)\\1{2,}").matcher(n);
   tempString = n;
    }
}
return n;
}
```

Figura 5: Función de contraer palabras.

Para comprobar que las modificaciones realizadas no tuvieron un efecto negativo,se decidió ejecutar las librerías con los mismos datos reportados por WEBIS y se comparó su F-score.

[1]https://github.com/dwyl/english-words

Tabla 1: Resultado librerías NRC.

Librerías	**Resultado**
WEBIS 2015	69,42 %
Experimento WEBIS 2016	69,24 %
NRC_UCSC	70,31 %
NRC-CANDAD 2013	69,29 %

WEBIS reporta un 69.44 % y experimentalmente se obtuvo un 69.24 % con la misma librería. Por otro lado, el resultado de experimentar con la librería modificada(NRC_UCSC) es de 70.31 %. Todos los resultados fueron superiores a los que consiguió el trabajo original NRC-CANADA 2013.

A pesar de que los resultados no son concluyentes, el hecho de realizar las modificaciones permitió generar y filtrar información.

4.3. Atributos propuestos

Como se buscó encontrar rasgos semánticos que enriquecieran al modelo de NRC-CANADA, cuyo enfoque se basa en rasgos de superficie junto con lexicones y partiendo de la premisa que la semántica apunta a aspectos de significado y sentido, se incorporaron Ngrams de sentidos, siguiendo la misma idea de generar atributos planteadas por este equipo. Debido a la sinonimia dos tweet semánticamente equivalentes, podrían contener Ngrams de palabras diferentes entre sí, mientras que los Ngrams de sentidos de ambos podrían ser los mismos.

Los sentidos o Synsets se recuperan de WordNet, utilizando la herramienta que

proporciona la librería de Python NLTK Bird et al. (2009) y esto se lleva a cabo de dos formas: la primera, es consultando el sentido más frecuente(Most Frecuent sense), donde se consulta a WordNet sin desambiguar, lo que quiere decir que no importara el contexto en el que se encuentre la palabra o token. Se entrega el primer sentido de la lista, dependiendo la etiqueta POS de la palabra a consultar.

Como el etiquetador POS de CMU pos-taggin tool utiliza etiquetas distintas a las de WordNet, se volvió a etiquetar el tweet con la herramienta disponible en la librería NLTK y se realizó una conversión de etiquetas para hacer compatible la consulta.

La segunda forma de obtención de sentidos, es desambiguando a través del algoritmo de Lesk, proporcionado por la librería de Python NLTK, quien para desambiguar un token utiliza la etiqueta POS de WordNet y el contexto, en este caso el tweet entero.

Como se extraen sentidos de WordNet, se puede consultar a SentiWordNet por la polaridad de los sentidos.

Se planteó la idea de estudiar la polaridad del último Emoticon. En el caso de los emoticones, se consideró el hecho de que cada uno de estos representa una carga semántica, que se puede asociar a una polaridad; por lo tanto, pueden existir dos emoticones distintos, pero que semánticamente sean iguales o que compartan la misma polaridad. Para generar este atributo, se utilizó el lexicón de emoticones,Emoticon Sentiment Lexicón (Hogenboom et al., 2015),el cual cuenta con 477 emoticones, donde su polaridad es representada por -1 (negativo), 0 (neutral) o 1 (positivo). Antes de consultar al lexicón, el Emoticon es Normalizado con la función de ContraerPalabra(Figura 5), con la opción de reducir el caracter repetido más de dos veces a uno. Por ejemplo, el Emoticon alargado ":))))))))" se convertirá en ":)", permitiendo un mejor uso del lexicón.

Considerando que las palabras alargadas podrían enfatizar un sentimiento expresado,se generó la opción de agregar y estudiar su polaridad .

Finalmente de estas características se proponen los siguientes atributos:

- Ngrams de sentido(NoWSDgram) : presencia o ausencia de secuencias contiguas de 1, 2, 3 y 4 tokens del sentido más frecuente.
- Ngrams de sentido(WSDgram): presencia o ausencia de secuencias contiguas de 1,2,3 y 4 tokens desambiguados con Lesk.
- Score polaridad SentiWordnet(SW): se evalúan las mismas características definidas en los lexicones de NRC-CANADA.
- Score polaridad palabras alargadas(Pol_ew): Se utilizan todos los lexicones del modelo NRC-CANADA y se evalúan las mismas características, descartando el análisis de bigrams y pares de palabras no contiguas.
- Polaridad del último Emoticon (Pol_le): Se analiza la presencia del Emoticon dentro del lexicón, Emoticon Sentiment Lexicón y se asigna su polaridad correspondiente, si no existe presencia de un último Emoticon como token o este no se encuentra en el lexicón, se le asigna un valor -2.

4.4. Etapas del sistema

Se generó un sistema, siguiendo el enfoque propuesto, donde se combinan los atributos del modelo NRC-CANADA y los propuesto, originando la posibilidad de crear más de 1 modelo. Como se adoptó un sistema base en la etapa de selección del modelo, sobre el mismo se trabajó y modificó. Por lo tanto, las mismas etapas que se mencionan en la descripción de la librería de WEBIS, son las descritas para el nuevo sistema.

4.4.1. Preprocesamiento

Esta etapa consiste en preparar los tweets, según las necesidades que presente el atributo.

Para extraer distintos atributos se requieren diferentes operaciones de preprocesamiento, según el tipo de rasgo, cuidando hacer sólo operaciones de preprocesamiento que requiere cada rasgo, además privilegiando la eficiencia al evitar repetir preprocesamiento común a varios rasgos.

Las operaciones de preprocesamiento que se consideran como esenciales para una correcta extracción de atributos, son las siguientes:

- Identificación de tweets repetidos.
- Conversión de todos los caracteres del abecedario dentro del tweet a minúscula.
- Reemplazo de nombres de usuarios y URL, por un espacio en blanco.
- Tokenizado y etiquetado POS a los tweets.
- Transformar los tweets a UTF-8.

Como la herramienta CMU pos-taggin tool abre los tweets en UTF-8, se aprovechó dicha característica para transformar los tweets en UTF-8.

Las siguientes operaciones de preproceamiento, dependerán de los atributos que se necesiten:

- Eliminación de stopwords.
- Obtención de sentidos.
- Obtención de polaridad de los sentidos en SentiWordNet.

Dichas operaciones de preprocesamiento, son generadas por la librería de Python NLTK.

La obtención de la polaridad de los sentidos dependerá de la generación previa de los sentidos; por lo tanto se pueden obtener los sentidos sin la polaridad, pero no así, la polaridad de los sentidos sin los sentidos.

4.4.2. Extracción de atributos

Los atributos se extraen de los flujos creados en el preprocesamiento y pueden ser de 3 tipos: los que son del tipo dinámico, estático y los que dependen del uso de lexicones.

Se considerarán atributos dinámicos a los que dependen del tweet,esto quiere decir, que según su contenido, se generarán los atributos. En esta clasificación se consideran:

- Ngrams.
- Chargrams.
- NoWSDgrams.
- WSDgrams.
- Cluster.
- Etiquetas POS.
- Emoticones.

Los atributos que aquí se extraigan serán únicos, a pesar que se repitan en otro tweet; esto quiere decir , si el Bigrams,"very good", está presente en más tweets, sólo se almacenará una vez en la lista de Ngrams, lo mismo ocurre en todos los atributos de esta clasificación.

Los atributos estáticos, no dependen del tweet. La extracción de estos atributos siempre está presente sin importar de que el tweet cuente con estas características, ya que son valores numéricos y su ausencia será representada con un 0. En esta clasificación se consideran:

- Palabras en mayúscula.
- Hashtag.
- Puntuación.
- Palabras alargadas.
- Negación.
- Emoticon (último Emoticon).

Los atributos que dependen de lexicones, también se pueden considerar como estáticos, pero se decidió diferenciarlos, debido a que no representan cantidad, si no que entregan un puntaje con respecto a la polaridad. En esta clasificación se consideran:

- Lexicones.
- SentiWordNet.
- Polaridad último Emoticon.
- Polaridad palabras alargadas.

En el caso de los emoticones, se considera un atributo dinámico y a la vez estático, debido a que este atributo considera la presencia o ausencia de los emoticones que se extraigan de los tweets(dinámico). Por ejemplo, si el corpus contiene 200 emoticones distintos, se analizará sobre esa cantidad; por otro lado , es un atributo estático, porque la consideración del último Emoticon está representada por un 0 cuando éste no está presente y un 1, en el caso contrario.

4.4.3. Etapa de aprendizaje o entrenamiento

Una vez concluida la etapa de extracción de atributos , se identifica cada tweet, con sus respectivos atributos. Ahora los tweets, serán representaciones vectoriales, lo que permitirá el entrenamiento del SVM, proporcionado por libLINEAR utilizado con sus valores por defecto.

4.4.4. Predicción o test

Para predecir los tweet, se repiten las etapas anteriormente vistas, con la diferencia que en la etapa de aprendizaje no se generará un modelo, sino que se utilizará el modelo ya entrenado para la predicción de polaridad.

Capítulo 5

5. Experimentos

Realizando un set de experimentos, se buscó la forma de comparar resultados, para conocer el desempeño de los modelos provenientes del sistema generado a partir del enfoque.

Con la herramienta estadística de Scipy (Librería perteneciente a Python), se realizó el test no paramétrico de Wilcoxon.

5.1. Selección de atributos

Se partió del hecho de que se cuenta con los atributos del modelo de NRC-CANADA; sin embargo, una de las características principales de dicho modelo, es el gran número de atributos que se generan en el sistema por su condición de dinámicos. Esto fue lo que motivó el interés de conocer la efectividad de dichos atributos y cómo se comportan si son enriquecidos con atributos semánticos.

No todos los atributos que se proponen en esta investigación son semánticos; el caso de la polaridad de la palabra alargada, es un atributo que cae en la clasificación de lexicones, pero también es un rasgo de superficie, al cual se le dio mayor importancia con el estudio de polaridad. Si bien se buscó enriquecer un modelo con rasgos semánticos, también se generó la posibilidad de agregar más peso a los rasgos de superficie con rasgos de lexicones, debido a que el enfoque postula la utilización de estos 3 rasgos.

Como se trabaja con sentidos y se planteó la utilización de 2 formas de extracción: desambiguar con Lesk (WSD Lesk)y el sentido más frecuente (MF sense), se estudió la efectividad de ambos.

El sistema es capaz de generar distintos modelos,por tanto, se identificó al conjunto de atributos provenientes de NRC-CANADA como (SB).

Para conseguir los resultados de las características a estudiar, se generó un set de 8 modelos SB prescindiendo de atributos dinámicos (Tabla 2). El símbolo ✓denota la presencia del atributo en el modelo, por lado × la ausencia del atributo.

Tabla 2: Set de modelos SB

Modelo	**Ngrams**	**Chargrams**	**Cluster**	**Alias**
1	✓	✓	✓	SB
2	✓	×	✓	SB_SCG
3	×	✓	✓	SB_SNG
4	×	×	✓	SB_SCG_SNG
5	✓	✓	×	SB_SCL
6	✓	×	×	SB_SCG_SCL
7	×	✓	×	SB_SNG_SCL
8	×	×	×	SB_SCG_SNG_SCL

Los atributos propuestos se agruparon y organizaron en 20 experimentos, como se muestra en la Tabla 3, donde Pol_ult_emo (A), Pol_p_alargada (B), Wsdgram (C), Nowsdgram (D), Sw(Wsdgram) (E) y Sw(Nowsdgram) (F). El exp0 se consideró, como el experimento de control. El símbolo ✓denota la presencia del atributo en el experimento, por lado × la ausencia del atributo.

Tabla 3: Combinación de atributos propuestos

	Experimentos																			
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
A	×	✓	×	×	×	×	×	✓	✓	✓	✓	✓	×	×	×	×	✓	✓	✓	✓
B	×	×	✓	×	×	×	×	✓	×	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓
C	×	×	×	✓	×	✓	×	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×
D	×	×	×	×	✓	×	✓	×	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
E	×	×	×	×	×	✓	×	×	×	×	✓	×	×	×	✓	×	×	×	✓	×
F	×	×	×	×	×	×	✓	×	×	×	×	✓	×	×	×	✓	×	×	×	✓

A cada modelo SB se le agregaron los Experimentos, resultando 8 librerías con 20 modelos distintos cada una, lo que dio un total de 160 combinaciones de atributos o modelos.

5.2. Set de datos

Se decidió utilizar los Corpus de tweets facilitados por SemEval entre los años 2013 al 2016 en la tarea de "Sentiment Analysis in Twitter", específicamente en la sub-tarea "Message Polarity Classification", debido a que estos se encuentran disponibles sin costo alguno. Otro punto a favor es que cuentan con Corpus de entrenamiento y test, además facilitan una herramienta para medir el F-score utilizando la misma escala de evaluación oficial. La estructura para ambos Corpus es la misma, contienen 2 ID para identificar el tweet, la etiqueta de polaridad, que para el caso del Corpus de entrenamiento varía entre "Positive", "Negative" y "Neutral", y para el test "UNKNOW" o "NA", dependiendo del año. El orden de la estructura se muestra en la Figura 6.

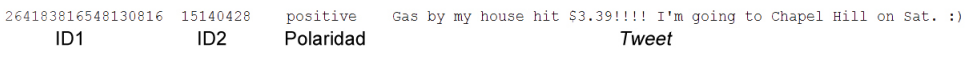

Figura 6: Estructura corpus de entrenamiento.

Para la realización de los experimentos 2013 al 2016, se utiliza el mismo corpus de entrenamiento(Tabla 4), sin embargo se filtran los tweet que no aportaron información significativa, lo cual esta permitido según las reglas de SemEval.

Tabla 4: Corpus entrenamiento oficial, con datos totales y filtrados.

Tweets	**Cantidad inicial**	**Cantidad final**
Positivos	4237	4215
Neutros	5339	5325
Negativos	1806	1708
Total	11382	11338

Pese a que se filtran algunos tweets, el Corpus posee tweets repetidos, lo cual no representa un problema, debido a que el sistema los detecta. Los corpus de test varían entre los años 2013 al 2016 (Tabla 5).

Tabla 5: Corpus de test, según su clasificación, entre los años 2013 al 2016.

Tweets	**Cantidad 2013**	**Cantidad 2014**	**Cantidad 2015**	**Cantidad 2016**
Positivos	1573	3506	1040	7059
Neutros	1640	3940	987	10342
Negativos	601	1541	365	3231
Total	3814	8987	2392	20632

5.3. Relación cantidad de atributos y F-score

Debido a que se postuló remover atributos pertenecientes a SB, que representan una gran cantidad de total extraído en el proceso de entrenamiento del SVM, se probó si con una menor cantidad de atributos se logran iguales o mejores resultados.

Calculando el F-score individual por modelo, evaluado en los Corpus de test entre los años 2013 al 2016, se determinó que no existe correlación, tras calcular el valor del coeficiente de determinación (R^2).

En el experimento de control, la librería SB_SCG, obtuvo los más altos F-score en todos los Corpus de test. Por otro lado la librería SB_SCG_SCL en los años 2014, 2015 y 2016, fue superior a la librería SB.

5.3.1. Relación cantidad de atributos y F-score Corpus 2013

Los datos de F-score calculados el año 2013 (Figura 7), presentan un valor R^2= 0,30. El equipo NRC-CANADA, ganador de SemEval para el 2013 obtuvo un 69,02 %, mientras que 114 modelos en esta investigación fueron superiores, donde mayor resultado obtenido es de un 71,33 % en la librería SB_SCG en el experimento 4 y el menor fue de un 65,24 % en la librería SB_SCG_SNG_SCL en el experimento 1 que comparado con el raking de SemEval ocuparía un tercer lugar.

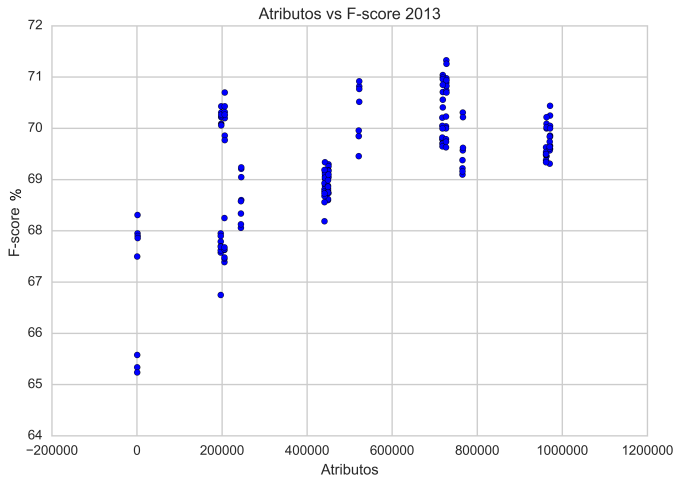

Figura 7: Gráfico de dispersión, Atributos vs F-score 2013.

5.3.2. Relación cantidad de atributos y F-score Corpus 2014

Los datos de F-score calculados el año 2014 (Figura 8), presentan un valor R^2= 0,0010. El equipo TEAMX, ganador de SemEval para el 2014 obtuvo un 70,96 %, mientras que 18 modelos en esta investigación fueron superiores, donde el mayor resultado obtenido es de un 72,50 % en la librería SB_SCG_SNG en el experimento 17 y el menor fue de un 63,77 % en la librería SB_SCG_SNG_SCL en el experimento 1 que comparado con el ranking de SemEval ocuparía un vigésimo segundo lugar.

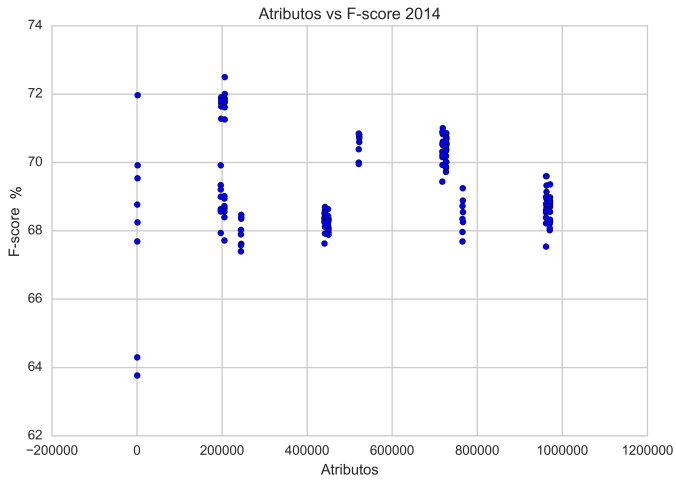

Figura 8: Gráfico de dispersión, Atributos vs F-score 2014.

5.3.3. Relación cantidad de atributos y F-score Corpus 2015

Los datos de F-score calculados el año 2015 (Figura 9), presentan un valor R^2= 0,2158. El equipo WEBIS, ganador de SemEval para el 2015 alcanzó un 64,84 %, mientras que 10 modelos en esta investigación fueron superiores, donde el resultado más alto obtenido es de un 66,24 % en la librería SB_SCG en el experimento 4 y el menor fue de un 58,46 % en la librería SB_SCG_SNG_SCL en el experimento 7 que comparado con el ranking de SemEval ocuparía un vigésimo lugar.

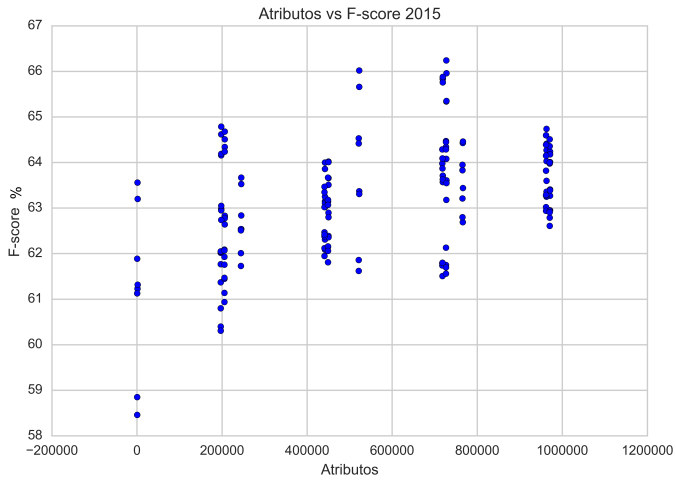

Figura 9: Gráfico de dispersión, Atributos vs F-score 2015.

5.3.4. Relación cantidad de atributos y F-score Corpus 2016

Los datos de F-score calculados el año 2016 (Figura 10), presentan un valor R^2= 0,27. El equipo Swiss-Cheese, ganador de SemEval para el 2016 alcanzó un 63,3 %, que fue superior a todos los obtenidos en esta investigación, donde el resultado mayor alcanzó a un 59,60 %, en la librería SB_SCG y el menor fue de un 55,30 % en la librería SB_SCG_SNG_SCL en el experimento 1 que comparado con el ranking de SemEval ocuparía un vigésimo lugar.

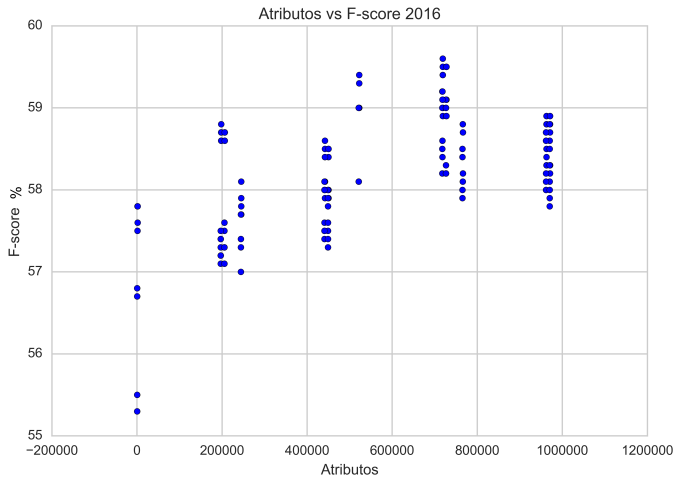

Figura 10: Gráfico de dispersión, Atributos vs F-score 2016.

Para todos los Corpus de test utilizados en esta investigación, el mayor resultado por año lo obtuvo un experimento del modelo SB, el cual no consideraba algún atributo dinámico. SB_SCG para los años 2013, 2015, 2016 y SB_SCG_SNG para el 2014. Se debe considerar que las librerías mencionadas trabajan con sólo el 32 % y 0,16 % de los atributos con respecto a la librería SB en el experimento de control. Dependiendo los experimentos ese porcentaje varía.

5.4. Efectividad Chargrams

Para conocer el efecto que tiene la extracción del atributo de Chargrams, se formaron 4 parejas de librerías, en las cuales sólo varía utilizar o prescindir del atributo Chargrams, como se muestra en la Tabla 6.

Tabla 6: Agrupación de parejas, para el análisis de la efectvidad Chargrams.

Pareja	**Librería con Chargrams**	**Librería sin Chargrams**
A	SB	SB_SCG
B	SB_SNG	SB_SCG_SNG
C	SB_SCL	SB_ SCG_SCL
D	SB_SNG_SCL	SB_SCG_SNG_SCL

Se considera que una pareja presentó un mayor rendimiento al prescindir de los Chargrams, si en más del 50 % de sus experimentos es superior a la librería que sí los utiliza. Se consideró la variable del número de atributos en caso de haberse presentado algún empate en el F-score; el que presentó menor cantidad de atributos resulto superior en la comparación.

Para determinar si existe diferencias estadísticas en la comparación de 2 librerías que conforman una pareja, se aplicó el test no paramétrico de los rangos signados de Wilcoxon, el cual propone que si el valor p es inferior a 0,05 existe diferencia significativa entre las medianas.

5.4.1. Efectividad Chargrams Corpus test SemEval 2013

En base a las parejas establecidas anteriormente, en la Figura 11, se puede observar la relación de prescindir o utilizar Chargrams con respecto a su F-score 2013. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza línea segmentada para las librerías sin Chargrams y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

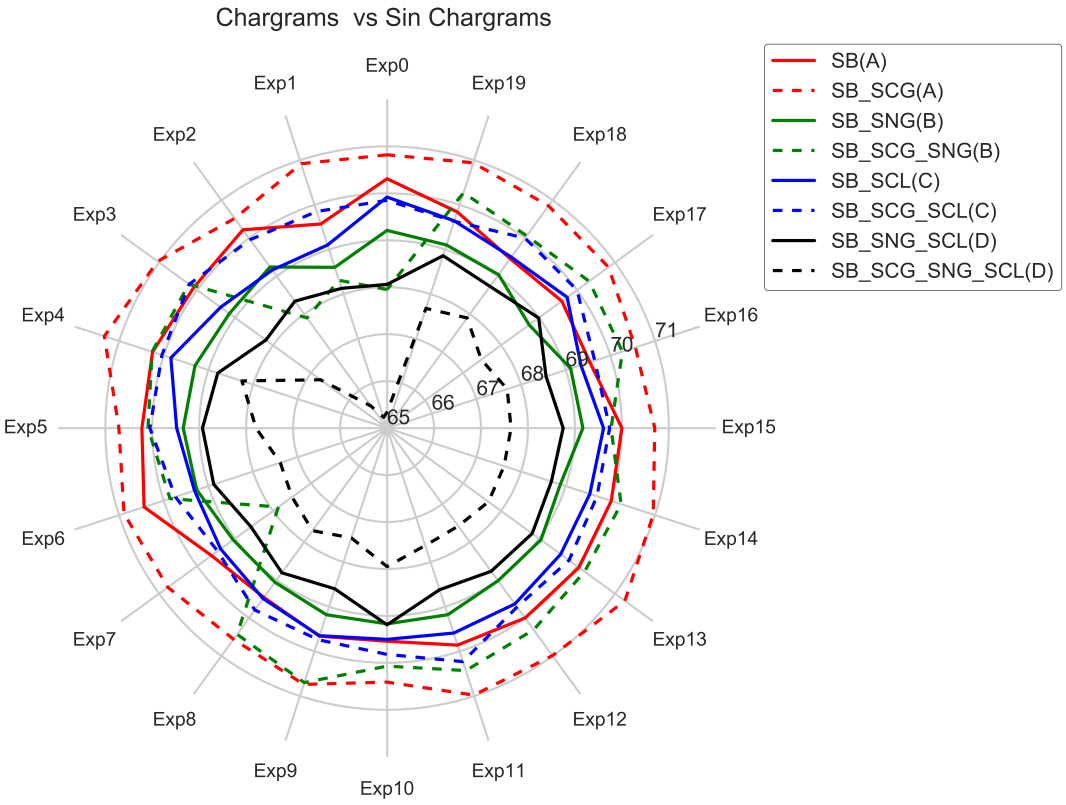

Figura 11: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Chargrams sobre el Corpus test SemEval 2013.

En 3 de las 4 parejas (A,B y C), el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Chargrams dentro de sus atributos, presenta un F-score más alto obtenido en el proceso de predicción de tweets, mientras que en la pareja restante (D), no sigue la tendencia de las anteriores, ya que el no uso de Chargrams disminuye sus resultados.

La Tabla 7 muestra la cantidad de experimentos en donde la librería sin Chargrams presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 7: Cantidad de pares de experimentos que presentan un mayor resultado sin Chargrams (se denota por CantExpSCG13) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2013.

Parejas	**CantExpSCG13**	**Valor P de Wilcoxon**
A	20	$8{,}84e^{-5}$
B	16	0,02
C	18	0,00013
D	0	$8{,}83e^{-5}$
Total	$(\frac{54}{80})100 = 67,50\,\%$	

Para el Corpus test SemEval 2013 el 67,50 % de los pares de experimentos aumenta sus resultados prescindiendo de los Chargrams. El mayor resultado obtenido por una librería sin Chargrams fue de 71,33 % en el experimento 4 en la librería SB_SCG de la pareja A, siendo este el mayor en la evaluación general. El resultado más bajo fue de un 65,24 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2013, resulto ser menor a un 0,05 lo que implica que existe una diferencia significativa entre las medias.

5.4.2. Efectividad Chargrams Corpus test SemEval 2014

En base a las parejas establecidas anteriormente, en la Figura 12, se puede observar la relación de prescindir o utilizar Chargrams con respecto a su F-score 2014. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza

línea segmentada para las librerías sin Chargrams y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

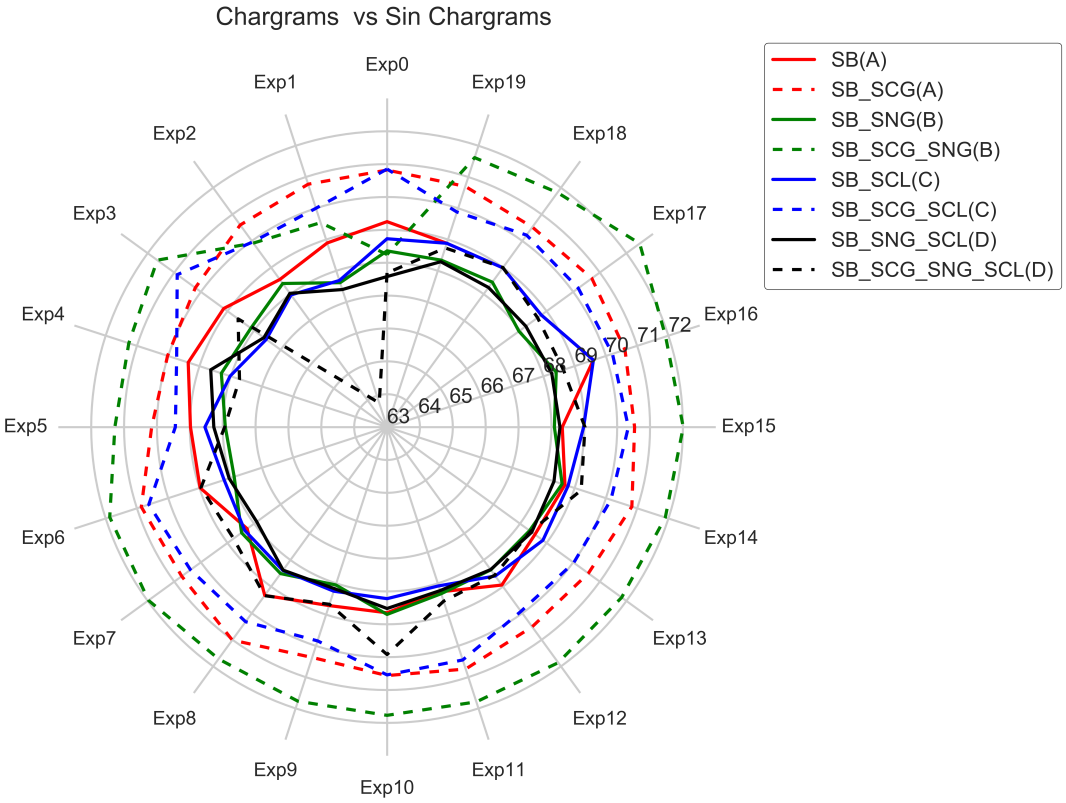

Figura 12: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Chargrams sobre el Corpus test SemEval 2014.

En todas las parejas, el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Chargrams dentro de sus atributos, presenta un F-score más alto obtenido en el proceso de predicción de tweets.

La Tabla 8 muestra la cantidad de experimentos en donde la librería sin Chargrams presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 8: Cantidad de pares de experimentos que presentan un mayor resultado sin Chargrams (se denota por CantExpSCG14) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2014.

Parejas	**CantExpSCG14**	**Valor P de Wilcoxon**
A	20	$8{,}84e^{-5}$
B	19	0,00010
C	20	$8{,}85e^{-5}$
D	15	0.085
Total	$(\frac{74}{80})100$=92,50 %	

Para el Corpus test SemEval 2014 el 92,50 % de los pares de experimentos aumenta sus resultados prescindiendo de los Chargrams. El mayor resultado obtenido por una librería sin Chargrams fue de 72,50 % en el experimento 17 en la librería SB_SCG_SNG de la pareja B, siendo este el mayor en la evaluación general. El resultado más bajo fue de un 63,27 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2014, resulto ser menor a un 0,05 en todas las parejas con excepción a la pareja D.

5.4.3. Efectividad Chargrams Corpus test SemEval 2015

En base a las parejas establecidas anteriormente, en la Figura 13, se puede observar la relación de prescindir o utilizar Chargrams con respecto a su F-score 2015. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza línea segmentada para las librerías sin Chargrams y en el caso contrario, línea continua,

mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

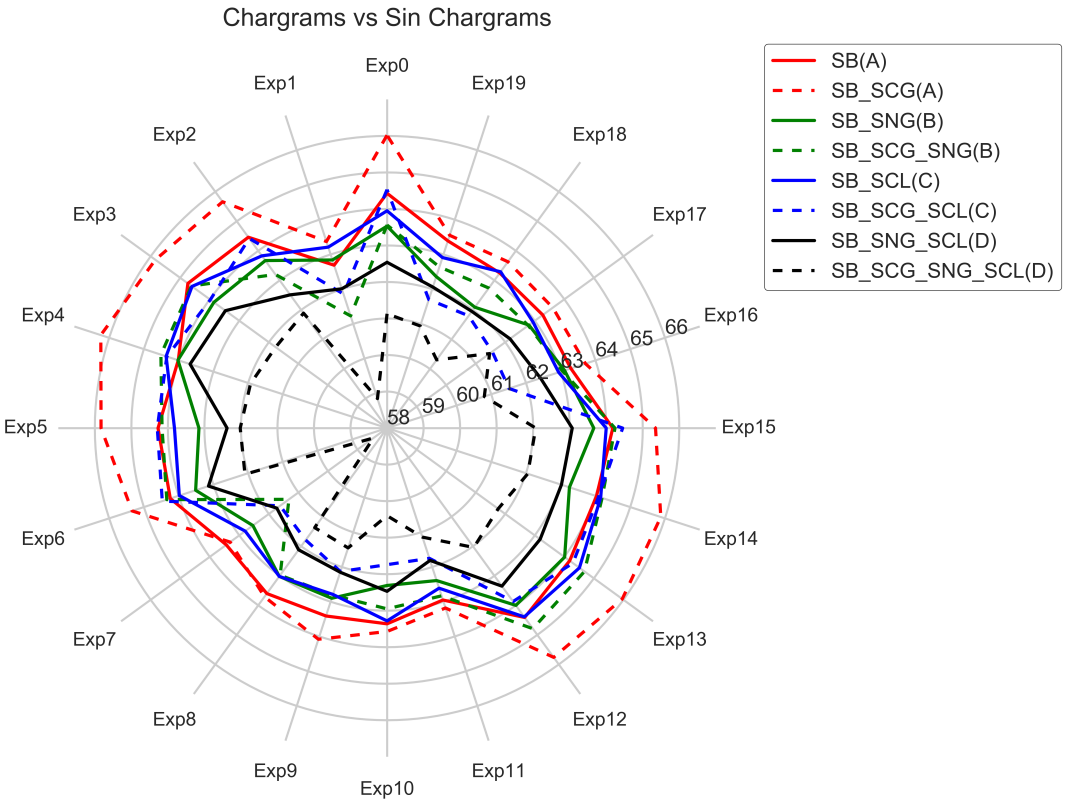

Figura 13: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Chargrams sobre el Corpus test SemEval 2015.

En 2 de las 4 parejas (A y B), el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Chargrams dentro de sus atributos, presenta un F-score más alto obtenido en el proceso de predicción de tweets, mientras que en las parejas restantes (C y D), no siguen la tendencia de las anteriores, ya que el no uso de Chargrams disminuye sus resultados. La Tabla9 muestra la cantidad de experimentos en donde la librería sin Chargrams presentó mayores resultados y el valor P de Wilcoxon

por pareja.

Tabla 9: Cantidad de pares de experimentos que presentan un mayor resultado sin Chargrams (se denota por CantExpSCG15) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2015.

Parejas	**CantExpSCG15**	**Valor P de Wilcoxon**
A	19	0,00012
B	13	0,092
C	5	0.0064
D	0	$8{,}85e^{-5}$
Total	$(\frac{37}{80})100$=46,20 %	

Para el Corpus test SemEval 2015 el 46,20 % de los pares de experimentos aumenta sus resultados prescindiendo de los Chargrams. El mayor resultado obtenido por una librería sin Chargrams fue de 66,24 % en el experimento 4 en la librería SB_SCG de la pareja A, siendo este el mayor en la evaluación general. El resultado más bajo fue de un 58,56 % de la librería SB_SCG_SNG_SCL en el experimento 7 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2015, resulto ser menor a un 0,05 en todas las parejas con excepción a la pareja B.

5.4.4. Efectividad Chargrams Corpus test SemEval 2016

En base a las parejas establecidas anteriormente, en la Figura 14, se puede observar la relación de prescindir o utilizar Chargrams con respecto a su F-score 2016. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza

línea segmentada para las librerías sin Chargrams y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

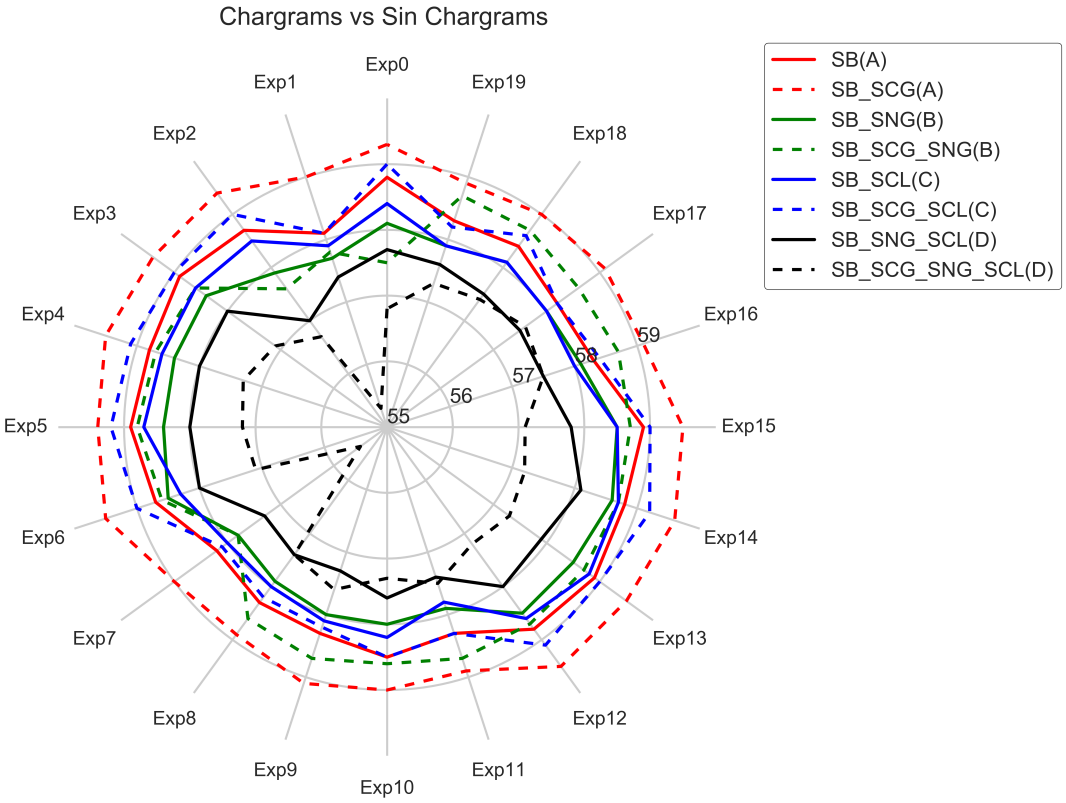

Figura 14: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Chargrams sobre el Corpus test SemEval 2016.

En 3 de las 4 parejas (A, B y C), el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Chargrams dentro de sus atributos, presenta un F-score más alto obtenido en el proceso de predicción de tweets, mientras que en las pareja restante (D), no sigue la tendencia de las anteriores, ya que el no uso de Chargrams disminuye sus resultados. La Tabla10 muestra la cantidad de experimentos

en donde la librería sin Chargrams presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 10: Cantidad de pares de experimentos que presentan un mayor resultado sin Chargrams (se denota por CantExpSCG16) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2016.

Parejas	**CantExpSCG16**	**Valor P de Wilcoxon**
A	20	$8{,}22e^{-5}$
B	18	0,0028
C	20	$7{,}71e^{-5}$
D	5	0,00099
Total	$(\frac{60}{80})100$=75,00 %	

Para el Corpus test SemEval 2016 el 75,00 % de los pares de experimentos aumenta sus resultados prescindiendo de los Chargrams. El mayor resultado obtenido por una librería sin Chargrams fue de 59,60 % en el experimento 14 en la librería SB_SCG de la pareja A, siendo este el mayor en la evaluación general. El resultado más bajo fue de un 55,30 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2016, resulto ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.4.5. Efectividad Chargrams global

En todos los Corpus test SemEval, el mayor resultado por Corpus lo obtuvo una librería sin presencia de Chargrams al igual que el más bajo. La Tabla 11, muestra el total individual y global por parejas, evaluadas en todos los Corpus utilizados.

Tabla 11: Cantidad de pares de experimentos que presentan un mayor resultado sin Chargrams sobre los Corpus test SemEval (se denota por AllCantExpSCG).

Parejas	**AllCantExpSCG**	**Total individual**
A	79	$(\frac{79}{80})100 = 98{,}75\,\%$
B	66	$(\frac{66}{80})100 = 82{,}50\,\%$
C	63	$(\frac{63}{80})100 = 78{,}75\,\%$
D	20	$(\frac{20}{80})100 = 25{,}00\,\%$
Total	$(\frac{228}{320})100 = 71{,}25\,\%$	

El 71,25 % de los pares de experimentos totales aumenta su resultado prescindiendo de los Chargrams. La librería que más se benefició de esta condición porcentualmente fue SB_SCG perteneciente a la pareja A; 3 de los 4 mejores resultados obtenidos de los Corpus test SemEval los obtuvo esta librería, seguida por SB_SCG_SNG, poseedora también de un primer lugar. Por otro lado, la pareja que más se vio afectada a nivel de resultados fue la pareja D, donde la librería SB_SCG_SNG_SCL fue la que obtuvo los 4 últimos lugares de los Corpus test SemEval.

El valor P de wilcoxon, para el Corpus de test SemEval 2013 y 2016,implica que la diferencia en las medianas de todas las parejas es significativa al usar o prescindir de Chargrams. Para los Corpus de test restantes se obtiene la misma conclusión en 3 de las 4 parejas.

5.5. Efectividad Ngrams

Para conocer el efecto que tiene la extracción del atributo de Ngrams, se formaron 4 parejas de librerías, en las cuales sólo varía utilizar o prescindir del atributo Ngrams, como se muestra en la Tabla 12

Tabla 12: Agrupación de parejas, para el análisis de la efectividad Ngrams.

Pareja	**Librería con Ngrams**	**Librería sin Ngrams**
A	**SB**	**SB_SNG**
B	**SB_SCG**	**SB_SCG_SNG**
C	**SB_SCL**	**SB_ SNG_SCL**
D	**SB_SCG_SCL**	**SB_SCG_SNG_SCL**

Se considera que una pareja presentó un mayor rendimiento al prescindir de los Ngrams, si en más del 50 % de sus experimentos es superior a la librería que sí los utiliza. Se consideró la variable del número de atributos en caso de haberse presentado algún empate en el F-score; el que presentó menor cantidad de atributos resultó superior en la comparación.

Para determinar si existe diferencias estadísticas en la comparación de 2 librerías que conforman una pareja , se aplicó el test no paramétrico de los rangos signados de Wilcoxon, el cual propone que si el valor p es inferior a 0,05 existe diferencia significativa entre las medianas.

5.5.1. Efectividad Ngrams Corpus test SemEval 2013

En base a las parejas establecidas anteriormente, en la Figura15, se puede observar la relación de prescindir o utilizar Ngrams con respecto a su F-score 2013. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza

línea segmentada para las librerías sin Ngrams y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

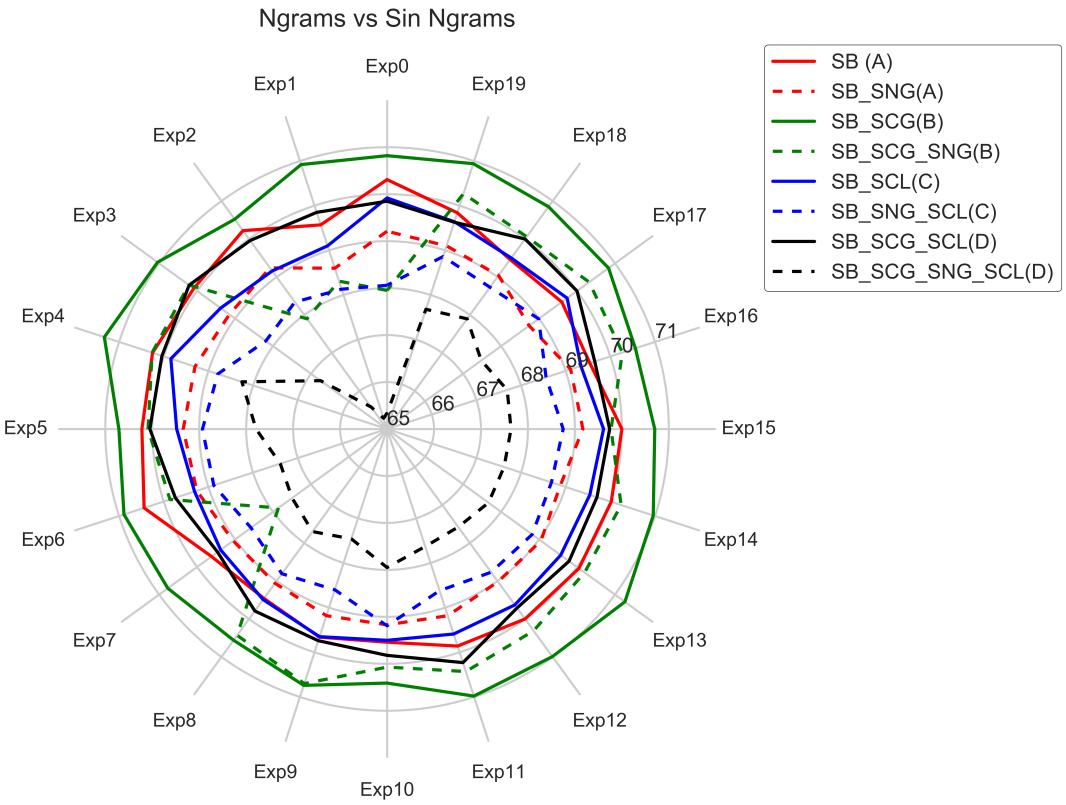

Figura 15: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Ngrams sobre el Corpus test SemEval 2013

En todas las parejas, el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Ngrams dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets.

La Tabla13 muestra la cantidad de experimentos en donde la librería sin Ngrams presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 13: Cantidad de pares de experimentos que presentan un mayor resultado sin Ngrams (se denota por CantExpSNG13) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2013.

Parejas	**CantExpSNG13**	**Valor P de Wilcoxon**
A	0	$8{,}84e^{-5}$
B	0	$8{,}85e^{-5}$
C	0	$8{,}76e^{-5}$
D	0	$8{,}84e^{-5}$
Total	$(\frac{0}{80})100$=0,00 %	

Para el Corpus test SemEval 2013 el 0 % de los pares de experimentos aumenta sus resultados prescindiendo de Ngrams. El mayor resultado obtenido por una librería sin Ngrams fue de 70,70 % en el experimento 9 en la librería SB_SCG_SNG de la pareja B, no siendo este el mayor en la evaluación general. El resultado más bajo fue de un 65,24 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2013, resulto ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.5.2. Efectividad Ngrams Corpus test SemEval 2014

En base a las parejas establecidas anteriormente, en la Figura16, se puede observar la relación de prescindir o utilizar Ngrams con respecto a su F-score 2014. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza

línea segmentada para las librerías sin Ngrams y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

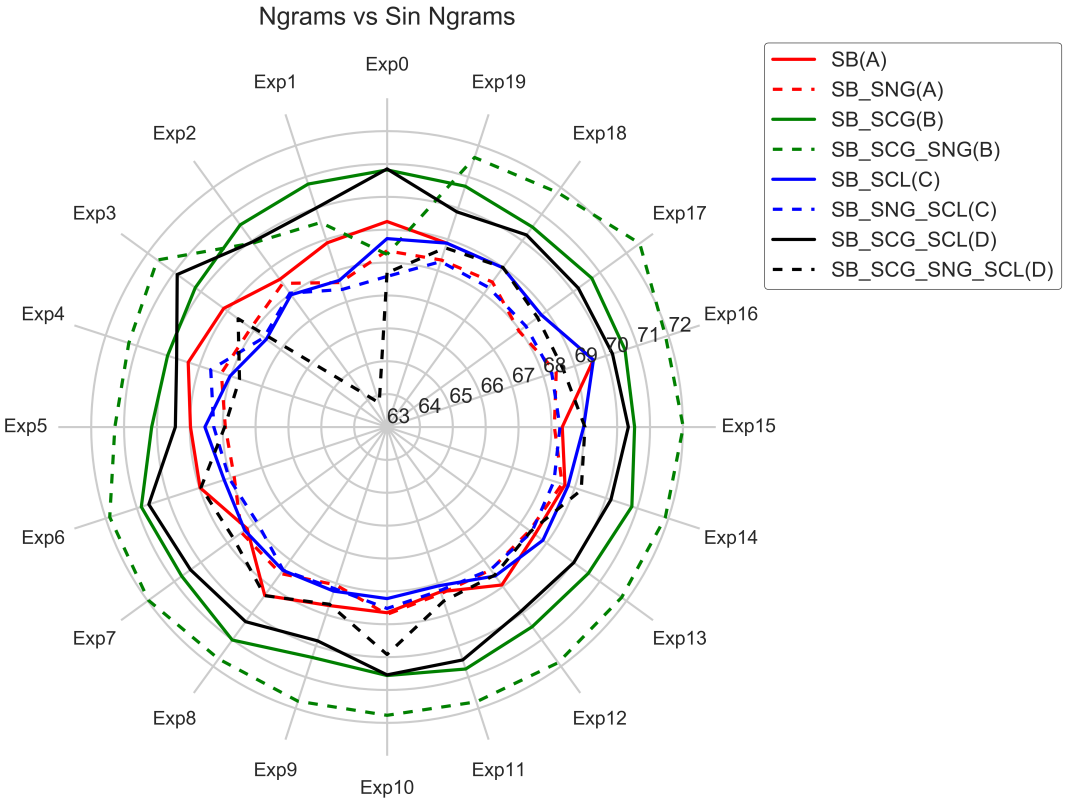

Figura 16: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Ngrams sobre el Corpus test SemEval 2014

En 3 de las 4 parejas (A, C y D), el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Ngrams dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets, mientras que en la pareja restante (B), no sigue la tendencia de las anteriores, ya que el no poseer Ngrams aumentó sus resultados.

La Tabla 14 muestra la cantidad de experimentos en donde la librería sin Ngrams presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 14: Cantidad de pares de experimentos que presentan un mayor resultado sin Ngrams (se denota por CantExpSNG14) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2014.

Parejas	**CantExpSNG14**	**Valor P de Wilcoxon**
A	3	0,00037
B	17	0,00510
C	5	0,0111
D	0	$8{,}85e^{-5}$
Total	$(\frac{25}{80})100$=31,25 %	

Para el Corpus test SemEval 2014 el 31,25 % de los pares de experimentos aumenta sus resultados prescindiendo de Ngrams. El mayor resultado obtenido por una librería sin Ngrams fue de 72,50 % en el experimento 17 en la librería SB_SCG_SNG de la pareja B, siendo el mayor en la evaluación general. El resultado más bajo fue de un 63,77 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2014, resultó ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.5.3. Efectividad Ngrams Corpus test SemEval 2015

En base a las parejas establecidas anteriormente, en la Figura17, se puede observar la relación de prescindir o utilizar Ngrams con respecto a su F-score 2015. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza línea segmentada para las librerías sin Ngrams y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

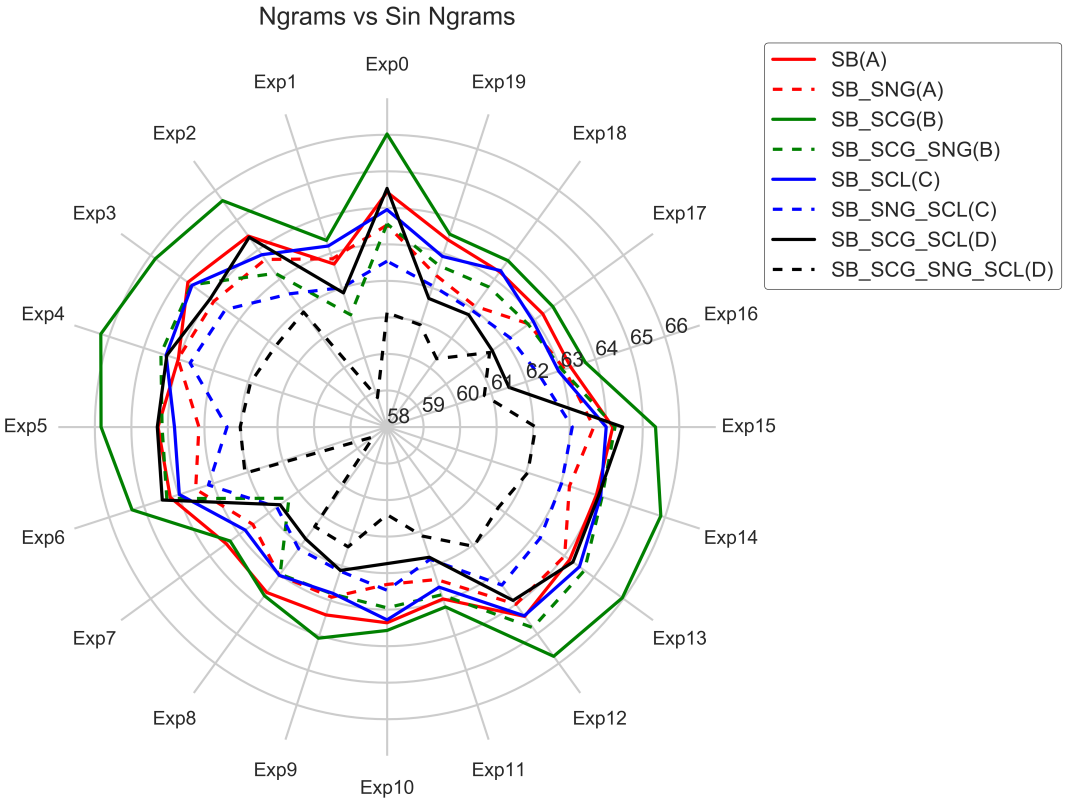

Figura 17: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Ngrams sobre el Corpus test SemEval 2015

De las 4 parejas, el rendimiento de más del 50,00 % de los experimentos en las

librerías que no poseen Ngrams dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets.

La Tabla15 muestra la cantidad de experimentos en donde la librería sin Ngrams presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 15: Cantidad de pares de experimentos que presentan un mayor resultado sin Ngrams (se denota por CantExpSNG15) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2015.

Parejas	**CantExpSNG15**	**Valor P de Wilcoxon**
A	0	0,00014
B	0	$8{,}84e^{-5}$
C	0	$8{,}85e^{-5}$
D	0	$8{,}85e^{-5}$
Total	$(\frac{0}{80})100$=0,00 %	

Para el Corpus test SemEval 2015 el 0,00 % de los pares de experimentos aumenta sus resultados prescindiendo de Ngrams. El mayor resultado obtenido por una librería sin Ngrams fue de 64,79 % en el experimento 12 en la librería SB_SCG_SNG de la pareja B, no siendo este el mayor en la evaluación general. El resultado más bajo fue de un 58,46 % de la librería SB_SCG_SNG_SCL en el experimento 7 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2015, resultó ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.5.4. Efectividad Ngrams Corpus test SemEval 2016

En base a las parejas establecidas anteriormente, en la Figura18, se puede observar la relación de prescindir o utilizar Ngrams con respecto a su F-score 2016. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza línea segmentada para las librerías sin Ngrams y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

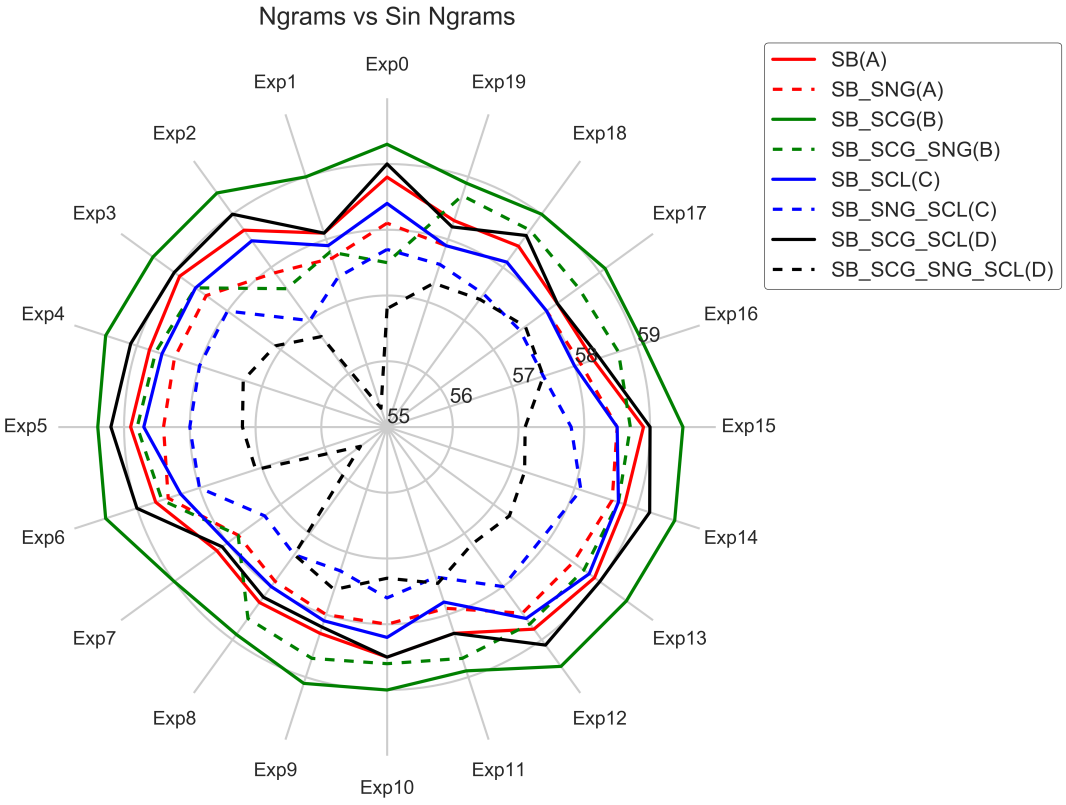

Figura 18: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Ngrams sobre el Corpus test SemEval 2016

De las 4 parejas, el rendimiento de más del 50,00 % de los experimentos en las

librerías que no poseen Ngrams dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets.

La Tabla16 muestra la cantidad de experimentos en donde la librería sin Ngrams presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 16: Cantidad de pares de experimentos que presentan un mayor resultado sin Ngrams (se denota por CantExpSNG16) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2016.

Parejas	**CantExpSNG16**	**Valor P de Wilcoxon**
A	0	$8{,}05e^{-5}$
B	0	$8{,}84e^{-5}$
C	0	$8{,}06e^{-5}$
D	0	$8{,}69e^{-5}$
Total	$(\frac{0}{80})100$=0,00 %	

Para el Corpus test SemEval 2016 el 0,00 % de los pares de experimentos aumenta sus resultados prescindiendo de Ngrams. El mayor resultado obtenido por una librería sin Ngrams fue de 58,80 % en el experimento 5 en la librería SB_SCG_SNG de la pareja B, no siendo este el mayor en la evaluación general. El resultado más bajo fue de un 55,30 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2016, resultó ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.5.5. Efectividad Ngrams global

En todos los Corpus test SemEval, el mayor resultado por Corpus lo obtuvo una librería que sí utiliza Ngrams, a excepción de lo ocurrido con el Corpus test SemEval 2014.No obstante a lo anterior, el resultado más bajo lo obtuvo una librería sin Ngrams. La Tabla 17, muestra el total individual y global por parejas, evaluadas en todos los Corpus utilizados.

Tabla 17: Cantidad de pares de experimentos que presentan un mayor resultado sin Ngrams sobre los Corpus test SemEval (se denota por AllCantExpSNG).

Parejas	**AllCantExpSNG**	**Total individual**
A	5	$(\frac{5}{80})100 = 6{,}25\,\%$
B	17	$(\frac{17}{80})100 = 21{,}25\,\%$
C	5	$(\frac{5}{80})100 = 6{,}25\,\%$
D	0	$(\frac{0}{80})100 = 0{,}00\,\%$
Total	$(\frac{27}{320})100 = 8{,}43\,\%$	

El 8,43 % de los pares de experimentos totales aumenta su resultado prescindiendo de los Ngrams. La librería que más se benefició de esta condición porcentualmente fue SB_SCG_SNG, perteneciente a la pareja B; 1 de los 4 mejores resultados obtenidos de los Corpus test SemEval los obtuvo esta librería. La pareja que más se vio afectada a nivel de resultados fue la pareja D, donde la librería SB_SCG_SNG_SCL fue la que obtuvo los 4 últimos lugares de los Corpus test SemEval.

El valor P de wilcoxon, para todos Corpus de test SemEval, implica que la diferencia en las medianas de todas las parejas es significativa al usar o prescindir de Ngrams.

5.6. Efectividad Clusters

Para conocer el efecto que tiene la extracción del atributo de Clusters, se formaron 4 parejas de librerías, en las cuales sólo varía utilizar o prescindir del atributo Cluster, como se muestra en la Tabla 18.

Tabla 18: Agrupación de parejas, para el análisis de la efectividad Clusters.

Pareja	**Librería con Clusters**	**Librería sin Clusters**
A	**SB**	**SB_SCL**
B	**SB_SCG**	**SB_SCG_SCL**
C	**SB_SNG**	**SB_ SNG_SCL**
D	**SB_SCG_SNG**	**SB_SCG_SNG_SCL**

Se considera que una pareja presentó un mayor rendimiento al prescindir de los Clusters, si en más del 50 % de sus experimentos es superior a la librería que sí los utiliza. Se consideró la variable del número de atributos en caso de haberse presentado algún empate en el F-score; el que presentó menor cantidad de atributos resulto superior en la comparación.

Para determinar si existe diferencias estadísticas en la comparación de 2 librerías que conforman una pareja , se aplicó el test no paramétrico de los rangos signados de Wilcoxon, el cual propone que si el valor p es inferior a 0,05 existe diferencia significativa entre las medianas.

5.6.1. Efectividad Cluster Corpus test SemEval 2013

En base a las parejas establecidas anteriormente, en la Figura19, se puede observar la relación de prescindir o utilizar Clusters con respecto a su F-score 2013. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza

línea segmentada para las librerías sin Clusters y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

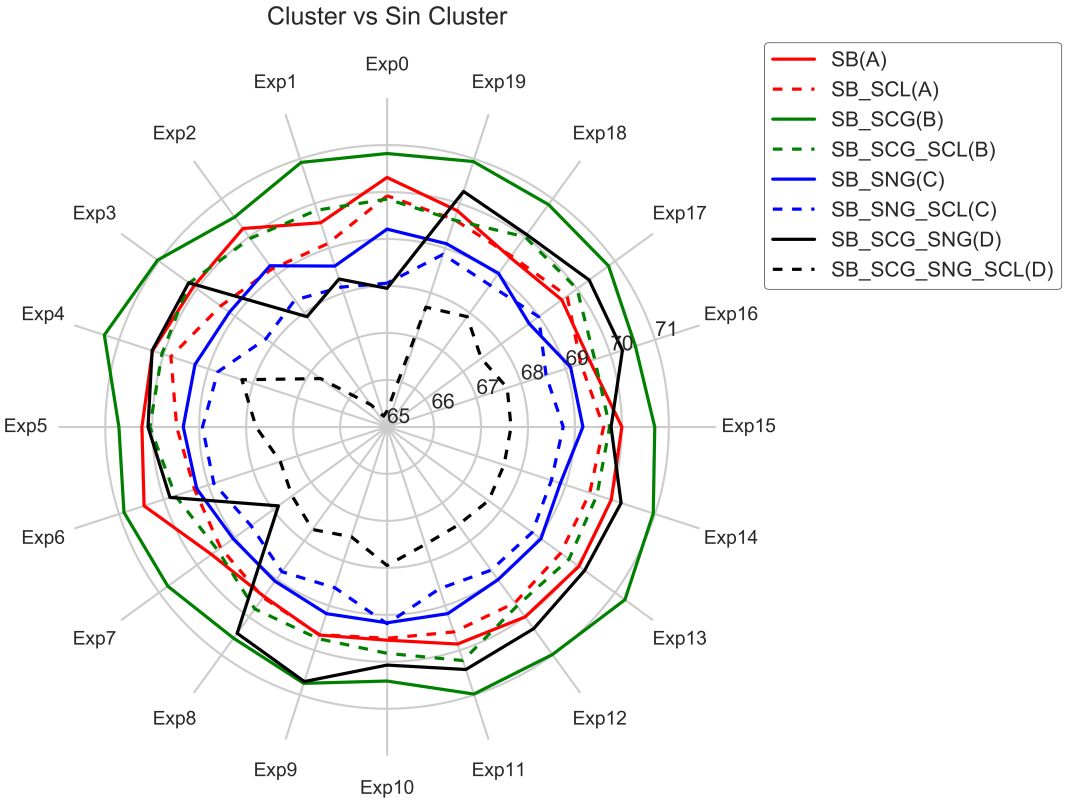

Figura 19: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Clusters sobre el Corpus test SemEval 2013

De las 4 parejas, el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Clusters dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets.

La Tabla19 muestra la cantidad de experimentos en donde la librería sin Clusters presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 19: Cantidad de pares de experimentos que presentan un mayor resultado sin Clusters (se denota por CantExpSCL13) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2013.

Parejas	**CantExpSCL13**	**Valor P de Wilcoxon**
A	3	0,0041
B	0	$8{,}84e^{-5}$
C	2	0,00025
D	0	$8{,}84e^{-5}$
Total	$(\frac{5}{80})100$=6,25 %	

Para el Corpus test SemEval 2013 el 6,25 % de los pares de experimentos aumenta sus resultados prescindiendo de Cluster. El mayor resultado obtenido por una librería sin Clusters fue de 70,04 % en el experimento 4 en la librería SB_SCG_SCL de la pareja B, no siendo este el mayor en la evaluación general. El resultado más bajo fue de un 65,24 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2013, resultó ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.6.2. Efectividad Cluster Corpus test SemEval 2014

En base a las parejas establecidas anteriormente, en la Figura20, se puede observar la relación de prescindir o utilizar Clusters con respecto a su F-score 2014. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza

línea segmentada para las librerías sin Clusters y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

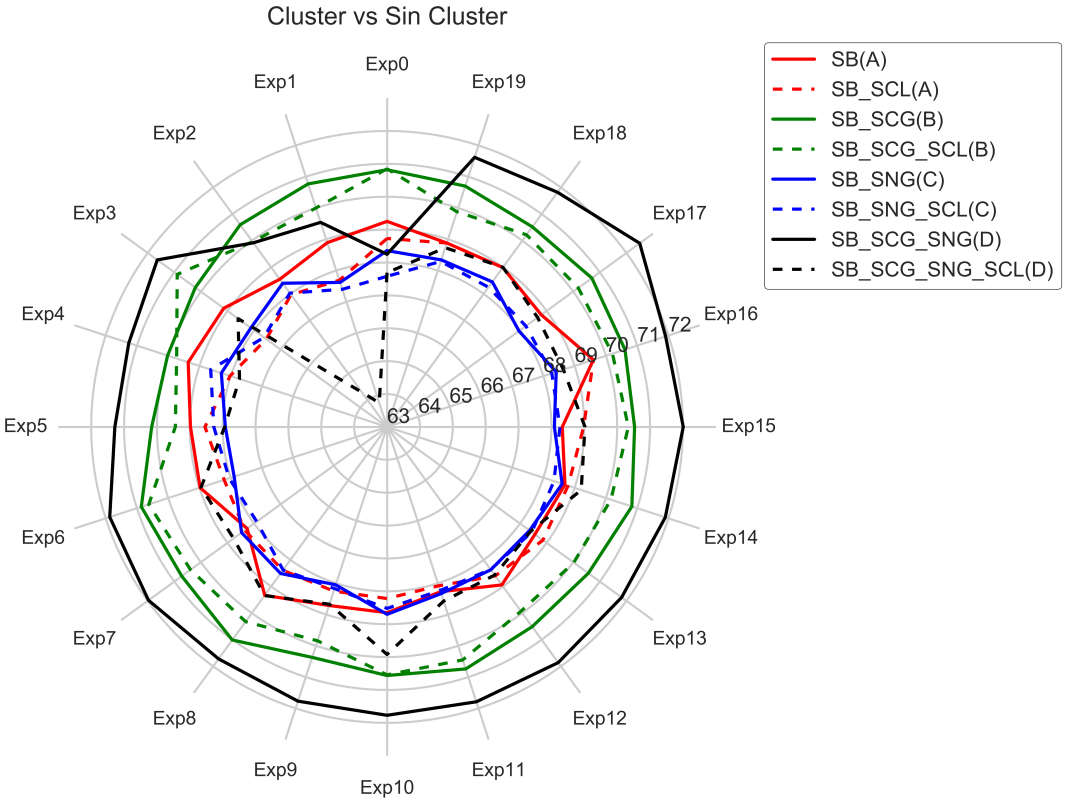

Figura 20: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Clusters sobre el Corpus test SemEval 2014

De las 4 parejas, el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Clusters dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets.

La Tabla20 muestra la cantidad de experimentos en donde la librería sin Clusters presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 20: Cantidad de pares de experimentos que presentan un mayor resultado sin Clusters (se denota por CantExpSCL14) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2014.

Parejas	**CantExpSCL14**	**Valor P de Wilcoxon**
A	8	0,0097
B	2	0,0011
C	7	0,19
D	0	$8,85e^{-5}$
Total	$(\frac{17}{80})100$=21,25 %	

Para el Corpus test SemEval 2014 el 21,25 % de los pares de experimentos aumenta sus resultados prescindiendo de Cluster. El mayor resultado obtenido por una librería sin Clusters fue de 70,89 % en el experimento 3 en la librería SB_SCG_SCL de la pareja B, no siendo éste el mayor en la evaluación general. El resultado más bajo fue de un 63,77 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2014, resultó ser menor a un 0,05 en todas las parejas con excepción a la pareja D.

5.6.3. Efectividad Cluster Corpus test SemEval 2015

En base a las parejas establecidas anteriormente, en la Figura21, se puede observar la relación de prescindir o utilizar Clusters con respecto a su F-score 2015. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza línea segmentada para las librerías sin Clusters y en el caso contrario, línea continua,

mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

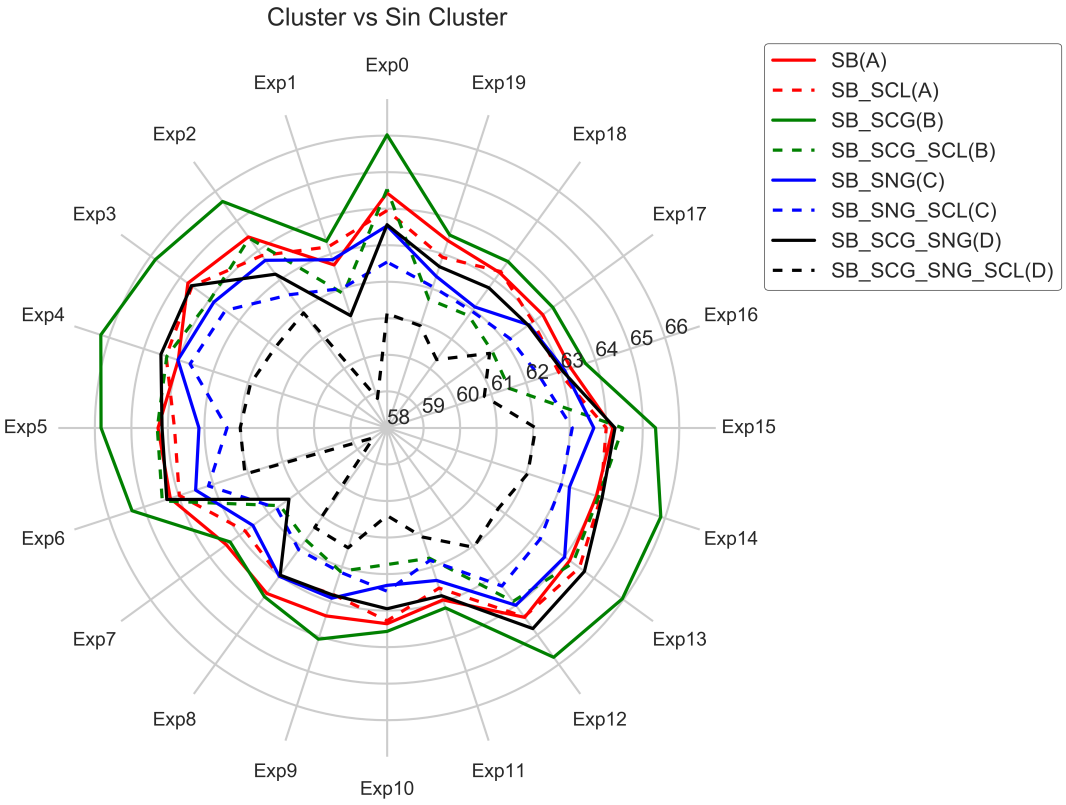

Figura 21: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Clusters sobre el Corpus test SemEval 2015

De las 4 parejas, el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Clusters dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets.

La Tabla21 muestra la cantidad de experimentos en donde la librería sin Clusters presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 21: Cantidad de pares de experimentos que presentan un mayor resultado sin Clusters (se denota por CantExpSCL15) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2015.

Parejas	**CantExpSCL15**	**Valor P de Wilcoxon**
A	5	0,016
B	0	$8{,}85e^{-5}$
C	1	0,00012
D	0	$8{,}81e^{-5}$
Total	$(\frac{6}{80})100$=7,50 %	

Para el Corpus test SemEval 2015 el 7,50 % de los pares de experimentos aumenta sus resultados prescindiendo de Cluster. El mayor resultado obtenido por una librería sin Clusters fue de 64,60 % en el experimento 4 en la librería SB_SCL de la pareja A, no siendo éste el mayor en la evaluación general. El resultado más bajo fue de un 58,46 % de la librería SB_SCG_SNG_SCL en el experimento 7 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2015, resultó ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.6.4. Efectividad Cluster Corpus test SemEval 2016

En base a las parejas establecidas anteriormente, en la Figura22, se puede observar la relación de prescindir o utilizar Clusters con respecto a su F-score 2016. Cada color representa a una pareja, donde se diferencian las librerías por el tipo de línea; se utiliza

línea segmentada para las librerías sin Clusters y en el caso contrario, línea continua, mientras más alejada del centro se encuentre el tipo de línea por Experimento (Exp), mejor es el resultado.

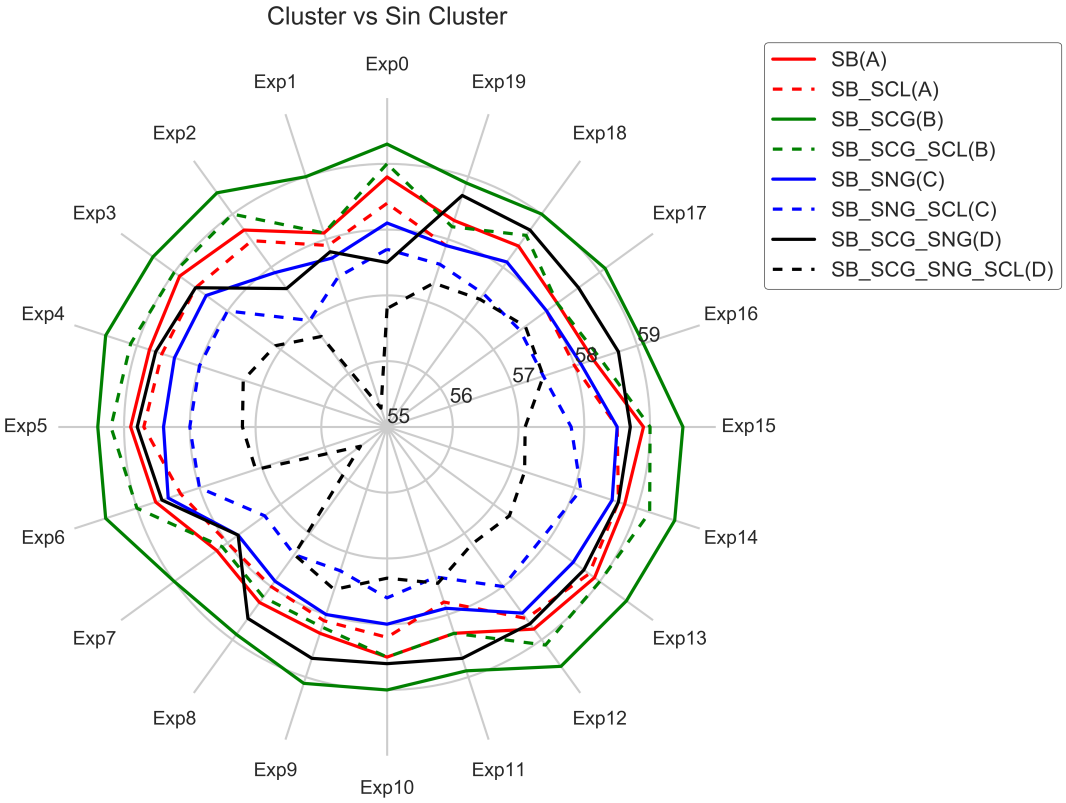

Figura 22: Gráfico de radar con resultados obtenidos por las parejas de librerías, en el análisis de la efectividad de Clusters sobre el Corpus test SemEval 2016

De las 4 parejas, el rendimiento de más del 50,00 % de los experimentos en las librerías que no poseen Clusters dentro de sus atributos, presenta un F-score más bajo obtenido en el proceso de predicción de tweets.

La Tabla22 muestra la cantidad de experimentos en donde la librería sin Clusters presentó mayores resultados y el valor P de Wilcoxon por pareja.

Tabla 22: Cantidad de pares de experimentos que presentan un mayor resultado sin Clusters (se denota por CantExpSCL16) y su valor P para el test de Wilconxon por pareja, sobre el Corpus test SemEval 2016.

Parejas	**CantExpSCL16**	**Valor P de Wilcoxon**
A	0	$8{,}31e^{-5}$
B	0	$8{,}17e^{-5}$
C	0	$7{,}88e^{-5}$
D	0	$8{,}62e^{-5}$
Total	$(\frac{0}{80})100$=0,00 %	

Para el Corpus test SemEval 2016 el 0,00 % de los pares de experimentos aumenta sus resultados prescindiendo de Cluster. El mayor resultado obtenido por una liberia sin Clusters fue de 59,20 % en el experimento 14 en la librería SB_SCG_SCL de la pareja B, no siendo éste el mayor en la evaluación general. El resultado más bajo fue de un 55,30 % de la librería SB_SCG_SNG_SCL en el experimento 1 de la pareja D, resultando ser el menor en la evaluación general.

El valor P de Wilcoxon evaluado en las parejas ejecutadas en el Corpus 2016, resultó ser menor a un 0,05 en todas las parejas, lo que implica que existe una diferencia significativa entre las medias.

5.6.5. Efectividad Cluster global

En todos los Corpus test SemEval, el mayor resultado por Corpus lo obtuvo una librería que sí utiliza Clusters. Por otro lado, el resultado más bajo lo obtuvo una librería sin Clusters. La Tabla 23, muestra el total individual y global por parejas, evaluadas

en todos los Corpus utilizados.

Tabla 23: Cantidad de pares de experimentos que presentan un mayor resultado sin Clusters sobre los Corpus test SemEval (se denota por AllCantExpSCL).

Parejas	**AllCantExpSCL**	**Total individual**
A	16	$(\frac{16}{80})100 = 20{,}00\,\%$
B	2	$(\frac{2}{80})100 = 2{,}50\,\%$
C	10	$(\frac{10}{80})100 = 12{,}05\,\%$
D	0	$(\frac{0}{80})100 = 0{,}00\,\%$
Total	$(\frac{28}{320})100 = 8{,}75\,\%$	

El 8,75 % de los pares de experimentos totales aumenta su resultado prescindiendo de los Clusters. La librería que más se benefició de esta condición porcentualmente fue SB_SCL, perteneciente a la pareja A; ninguno de los mejores resultados obtenidos de los Corpus test SemEval los obtuvo esta librería. La pareja que más se vio afectada a nivel de resultados fue la pareja D, donde la librería SB_SCG_SNG_SCL fue la que obtuvo los 4 últimos lugares de los Corpus test SemEval.

El valor P de wilcoxon, para el Corpus de test SemEval 2013, 2015 y 2016, implica que la diferencia en las medianas de todas las parejas es significativa al usar o prescindir de Clusters. Para el Corpus de test restante se obtiene la misma conclusión en 3 de las 4 parejas.

5.7. Efectividad de Lesk vs Sentido más frecuente

Para conocer el efecto que tiene desambiguar el tweet, en el proceso de extracción de sentidos, se agruparon los experimentos que utilizan sentidos y se clasificaron según

su método de obtención. Esto quiere decir, por un lado se dejan fijo los atributos que no dependan de sentidos y sólo varía el hecho de agregar los sentidos obtenidos por el método de desambiguación de Lesk(WSD Lesk) o por el Sentido más frecuente(MF sense), generando 8 grupos de atributos (Tabla 24). Se considera que una pareja presentó un mayor rendimiento al utilizar Lesk o MF frequent sense, si en más del 50 % de sus experimentos es superior a la otra. Se consideró la variable del número de atributos en caso de haberse presentado algún empate en el F-score; el que presentó menor cantidad de atributos resulto superior en la comparación.

Para la verificación de efectividad de Lesk vs Most Frequent sense, no se aplicó el test no paramétrico de los rangos signados de Wilcoxon, por el hecho de que cada grupo de atributos, sólo aportó con 8 pares de datos, lo que es menos del 50 % del total de datos recomendados por la herramienta estadística de Scipy, con la cual se efectuaron los cálculos.

Tabla 24: Grupos de atributos fijos donde varía el método de desambiguación.

Atributos	**Método**	**Grupo**	**Experimentos**
-	WSD Lesk	1	Exp3
	MF sense		Exp4
Sw	WSD Lesk	2	Exp5
	MF sense		Exp6
Pol_le	WSD Lesk	3	Exp8
	MF sense		Exp9
Pol_le + Sw	WSD Lesk	4	Exp10
	MF sense		Exp11
Pol_ew	WSD Lesk	5	Exp12
	MF sense		Exp13
Pol_ew +Sw	WSD Lesk	6	Exp14
	MF sense		Exp15
Pol_le + Pol_ew	WSD Lesk	7	Exp16
	MF sense		Exp17
Pol_le + Pol_ew Sw	WSD Lesk	8	Exp18
	MF sense		Exp19

Para efectos de análisis a las librerías SB se les agrega un nuevo alias(Tabla25).

Tabla 25: Nuevos identificadores de librerías SB, para la efectividad de Lesk vs Most Frequent.

Librería	**Alias**
SB	L1
SB_SCG	L2
SB_SNG	L3
SB_SCG_SNG	L4
SB_SCL	L5
SB_SCG_SCL	L6
SB_SNG_SCL	L7
SB_SCG_SNG_SCL	L8

Cada librería ejecutará individualmente a todos los grupos clasificados anteriormente, generando 8 pares de combinaciones librería-grupo.

5.7.1. Efectividad de Lesk vs Most Frequent Corpus test SemEval 2013

En base a los grupos y librerías establecidas anteriormente, en la Figura 23, se puede observar la relación de utilizar Lesk o MF sense con respecto a su F-score 2013. Cada color representa a un método, donde se diferencian por el tipo de línea; se utiliza línea segmentada para MF sense y para Lesk, línea continua, mientras más alejada del centro se encuentre el tipo de línea por librería y grupo, mejor es el resultado. La nomenclatura que se utilizó en los ejes corresponde a Librería-grupo.

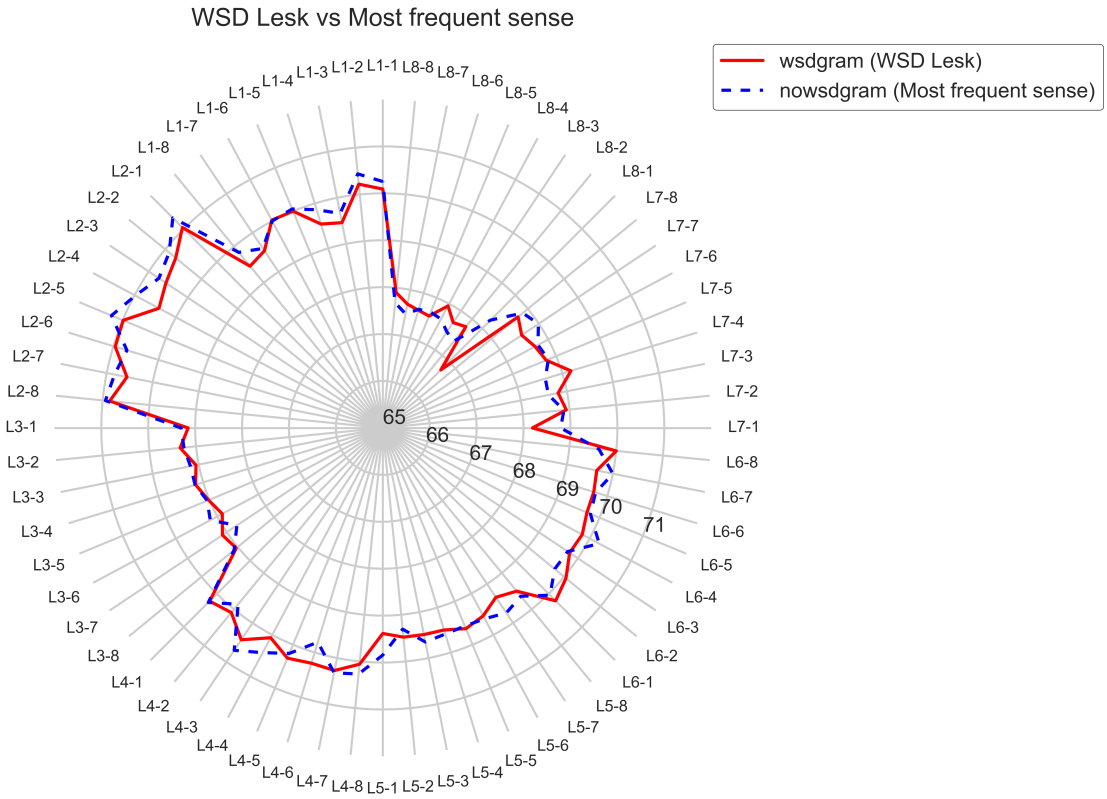

Figura 23: Gráfico de radar con resultados obtenidos por las combinaciones librería-grupo, en el análisis de la efectividad de Lesk vs Most frequent sense,sobre el Corpus test SemEval 2013.

Sólo en el grupo 2, el rendimiento de más del 50 % de las combinaciones librería-grupo que utilizan el algoritmo de Lesk para desambiguar, presenta mayores resultados en el F-score obtenido en el proceso de predicción de tweets, mientras que en las combinaciones restantes su uso empeora los resultados.

La Tabla 26 muestra la cantidad de combinaciones librería-grupo, donde el algoritmo de Lesk fue superior en el F-score.

Tabla 26: Cantidad de combinaciones librería-grupo que presentan un mayor resultado desambiguando con Lesk (se denota por CantExpLesk13), sobre el Corpus SemEval 2013.

Parejas	**CantExpLesk13**
Grupo 1	1
Grupo 2	6
Grupo 3	3
Grupo 4	2
Grupo 5	2
Grupo 6	3
Grupo 7	2
Grupo 8	2
Total	$(\frac{21}{64})100 = 32{,}81\,\%$

Para el Corpus test SemEval 2013 el 32,81 % de las combinaciones librería-grupo aumentan sus resultados utilizando Lesk. El mayor resultado obtenido usando Lesk fue de 71,04 % en el Grupo 1, experimento 3, en la librería L2, no siendo éste el mayor en la evaluación general del método. El resultado más bajo fue de un 66,75 % en el grupo 1, experimento 3, en la librería 8, resultando ser el menor en la evaluación general del método.

5.7.2. Efectividad de Lesk vs Most Frequent Corpus test SemEval 2014

En base a los grupos y librerías establecidas anteriormente, en la Figura 24, se puede observar la relación de utilizar Lesk o MF sense con respecto a su F-score 2014. Cada color representa a un método, donde se diferencian por el tipo de línea; se utiliza

línea segmentada para MF sense y para Lesk, línea continua, mientras más alejada del centro se encuentre el tipo de línea por librería y grupo, mejor es el resultado. La nomenclatura que se utilizó en los ejes corresponde a Librería-grupo.

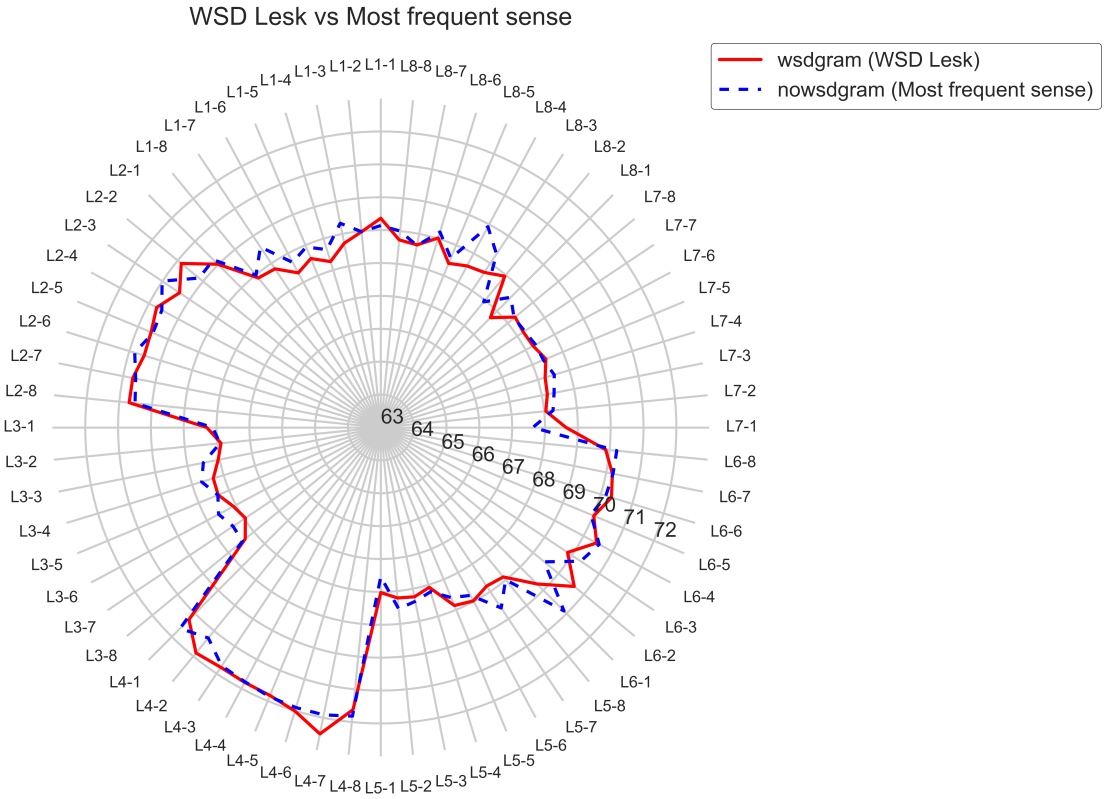

Figura 24: Gráfico de radar con resultados obtenidos por las combinaciones librería-grupo, en el análisis de la efectividad de Lesk vs Most Frequent sense, sobre el Corpus test SemEval 2014.

En 5 grupos(3,4,6,7 y 8), el rendimiento de más del 50 % de las combinaciones librería-grupo que utilizan el algoritmo de Lesk para desambiguar, presenta mayores resultados en el F-score obtenido en el proceso de predicción de tweets, mientras que en las combinaciones restantes su uso empeora los resultados.

La Tabla 27 muestra la cantidad individual por grupo y total de combinaciones librería-grupo, donde el algoritmo de Lesk fue superior en el F-score.

Tabla 27: Cantidad de combinaciones librería-grupo que presentan un mayor resultado desambiguando con Lesk (se denota por CantExpLesk14), sobre el Corpus SemEval 2014.

Parejas	**CantExpLesk14**
Grupo 1	4
Grupo 2	4
Grupo 3	7
Grupo 4	6
Grupo 5	4
Grupo 6	5
Grupo 7	5
Grupo 8	6
Total	$(\frac{41}{64})100 = 64{,}06\,\%$

Para el Corpus test SemEval 2014 el 64,06 % de las combinaciones librería-grupo aumentan sus resultados utilizando Lesk. El mayor resultado obtenido usando Lesk fue de 71,91 % en el Grupo 7, experimento 16, en la librería L4, no siendo éste el mayor en la evaluación general del método. El resultado más bajo fue de un 67,54 % en el grupo 1, experimento 3, en la librería 5, resultando ser el menor en la evaluación general del método.

5.7.3. Efectividad de Lesk vs Most Frequent Corpus test SemEval 2015

En base a los grupos y librerías establecidas anteriormente, en la Figura 25, se puede observar la relación de utilizar Lesk o MF sense con respecto a su F-score 2015. Cada color representa a un método, donde se diferencian por el tipo de línea; se utiliza línea segmentada para MF sense y para Lesk, línea continua, mientras más alejada del centro se encuentre el tipo de línea por librería y grupo, mejor es el resultado. La nomenclatura que se utilizó en los ejes corresponde a Librería-grupo.

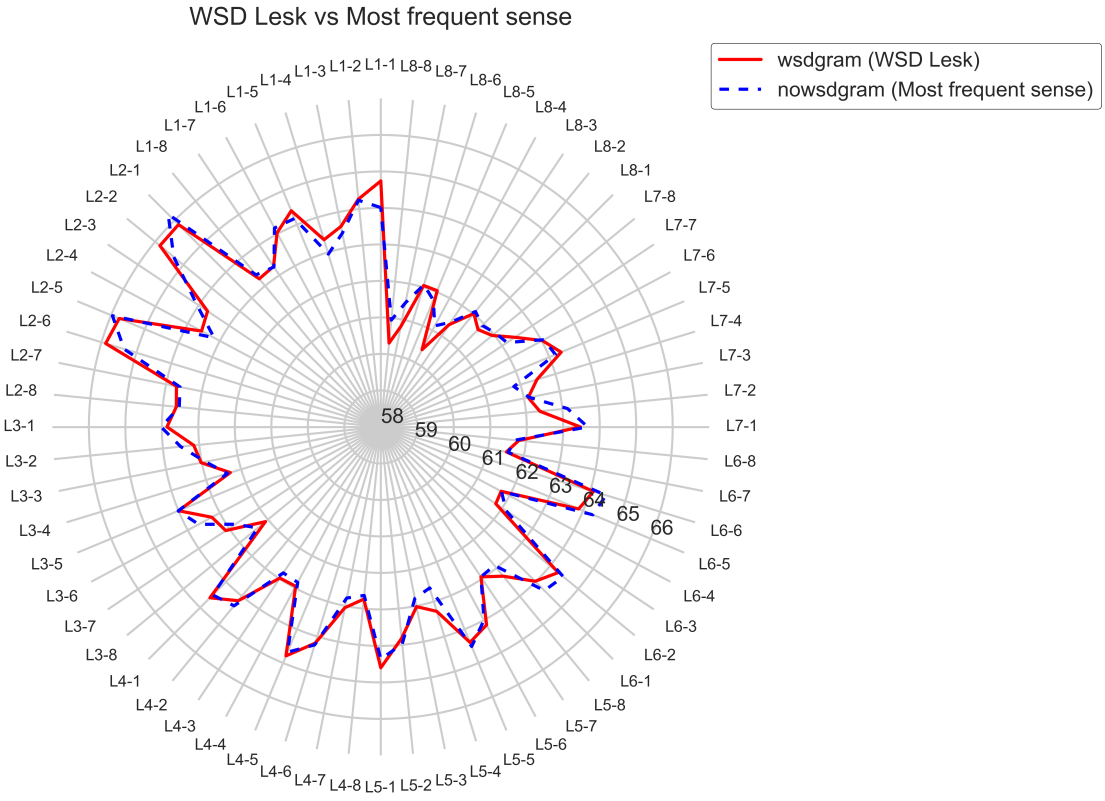

Figura 25: Gráfico de radar con resultados obtenidos por las combinaciones librería-grupo, en el análisis de la efectividad de la efectividad de Lesk vs Most frequent sense, sobre el Corpus test SemEval 2015.

En 2 grupos(4 y 7), el rendimiento de más del 50 % de las combinaciones librería-grupo que utilizan el algoritmo de Lesk para desambiguar, presenta mayores resultados en el F-score obtenido en el proceso de predicción de tweets, mientras que las combinaciones restantes su uso empeora los resultados.

La Tabla 28 muestra la cantidad la cantidad de combinaciones librería-grupo, donde el algoritmo de Lesk fue superior en el F-score.

Tabla 28: Cantidad de combinaciones librería-grupo que presentan un mayor resultado desambiguando con Lesk (se denota por CantExpLesk15), sobre el Corpus SemEval 2015.

Parejas	**CantExpLesk15**
Grupo 1	3
Grupo 2	2
Grupo 3	4
Grupo 4	5
Grupo 5	4
Grupo 6	2
Grupo 7	6
Grupo 8	4
Total	$(\frac{30}{64})100 = 46{,}87\,\%$

Para el Corpus test SemEval 2015 el 46,87 % de las combinaciones librería-grupo aumentan sus resultados utilizando Lesk. El mayor resultado obtenido usando Lesk fue de 65,88 % en el Grupo 6, experimento 14, en la librería L2, no siendo éste el mayor en la evaluación general del método. El resultado más bajo fue de un 60,31 % en el grupo 8, experimento 18, en la librería 8, resultando ser el menor en la evaluación general del

método.

5.7.4. Efectividad de Lesk vs Most frequent Corpus test SemEval 2016

En base a los grupos y librerías establecidas anteriormente, en la Figura 26, se puede observar la relación de utilizar Lesk o MF sense con respecto a su F-score 2016. Cada color representa a un método, donde se diferencian por el tipo de línea; se utiliza línea segmentada para MF sense y para Lesk, línea continua, mientras más alejada del centro se encuentre el tipo de línea por librería y grupo, mejor es el resultado. La nomenclatura que se utilizo en los ejes corresponde a Librería-grupo.

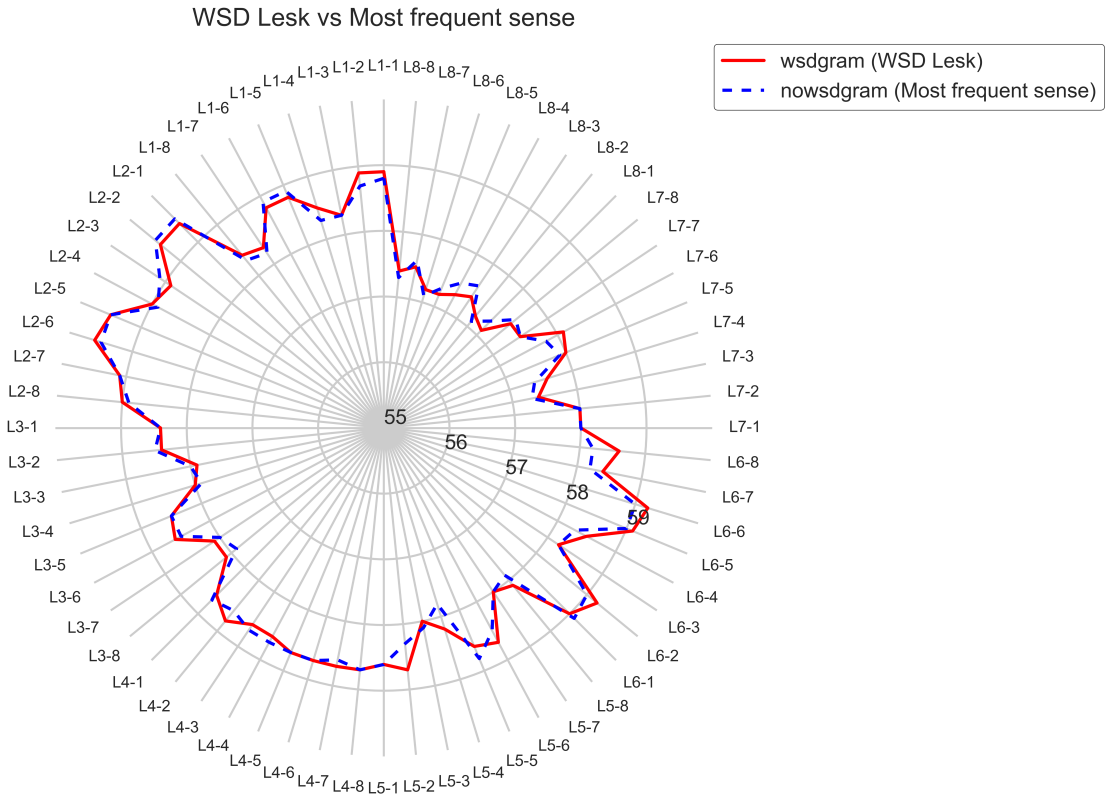

Figura 26: Gráfico de radar con resultados obtenidos por las combinaciones de librería-grupo, en el análisis de la efectividad de Lesk vs Most frequent sense, sobre el Corpus test SemEval 2016.

En 6 grupos(2,4,5,6,7 y 8), el rendimiento de más del 50 % de las combinaciones librería-grupo que utilizan el algoritmo de Lesk para desambiguar, presenta mayores resultados en el F-score obtenido en el proceso de predicción de tweets, mientras que en las combinaciones restantes su uso empeora los resultados.

La Tabla 29 muestra la cantidad de combinaciones librería-grupo, donde el algoritmo de Lesk fue superior en el F-score.

Tabla 29: Cantidad de combinaciones librería-grupo que presentan un mayor resultado desambiguando con Lesk(se denota por CantExpLesk16), sobre el Corpus SemEval 2016.

Parejas	**CantExpLesk16**
Grupo 1	4
Grupo 2	5
Grupo 3	3
Grupo 4	6
Grupo 5	5
Grupo 6	6
Grupo 7	7
Grupo 8	7
Total	$(\frac{43}{64})100 = 67{,}18\,\%$

Para el Corpus test SemEval 2016 el 67,18 % de las combinaciones librería-grupo aumentan sus resultados utilizando Lesk. El mayor resultado obtenido usando Lesk fue de 59,60 % en el Grupo 6, experimento 14, en la librería L2, siendo éste el mayor en la evaluación general del método. El resultado más bajo fue de un 57,10 % en el grupo 1, experimento 3, en la librería 8, no fue el menor dentro de esta evaluación por un desempate a nivel de atributos.

5.7.5. Efectividad de Lesk vs Most frequent sense global

En todos los Corpus test SemEval, el mayor resultado por Corpus lo obtuvo una combinación librería-grupo que no utiliza Lesk como método de desambiguacion, a excepción del Corpus test SemEval 2016,la misma situación se presentó con el menor

F-score. La Tabla 30, muestra el total individual y global por combinaciones librería-grupo, evaluadas en todos los Corpus utilizados.

Tabla 30: Cantidad de combinaciones librería-grupo que presentan un mayor resultado desambiguando con Lesk (se denota por AllCantExpLesk), sobre los Corpus SemEval.

Parejas	**AllCantExpLesk**	**Total individual**
Grupo 1	12	$(\frac{12}{32})100 = 37{,}50\,\%$
Grupo 2	17	$(\frac{17}{32})100 = 53{,}12\,\%$
Grupo 3	17	$(\frac{17}{32})100 = 53{,}12\,\%$
Grupo 4	19	$(\frac{19}{32})100 = 59{,}37\,\%$
Grupo 5	15	$(\frac{15}{32})100 = 46{,}87\,\%$
Grupo 6	16	$(\frac{16}{32})100 = 50{,}00\,\%$
Grupo 7	20	$(\frac{20}{32})100 = 62{,}50\,\%$
Grupo 8	19	$(\frac{19}{32})100 = 59{,}37\,\%$
Total	$(\frac{135}{256})100 = 52{,}73\,\%$	

El 52,73 % de las combinaciones librería-grupo totales, aumentan su resultado al utilizar Lesk. El grupo de atributos que más se beneficio por esta condición porcentualmente fue el grupo 7, mientras el grupo 1 fue el menos beneficiado.

Los resultado más bajos obtenidos en esta experimentación, no fueron los más bajos si se consideran los 160 modelos que se generan por corpus. Por otro lado, todos los F-score obtenidos fueron fueron los más altos.

Capítulo 6

6. Conclusión y trabajo futuro

Se logró conocer lo distintos enfoques utilizados para el análisis de polaridad en twitter, a través de la revisión bibliográfica en distintas fuentes científicas, lo que permitió además manejar conceptos claves y conocer de herramientas dedicadas al estudio de la red social. Se definió un sistema para la asignación de polaridad, donde se adoptó un modelo que combinara rasgos de superficie y lexicones, el que luego fue enriquecido con atributos semánticos y una mezcla entre rasgos de superficie y de lexicones. Para validar los modelos creados a partir del enfoque propuesto se creó un set de 160 modelos para realizar pruebas comparativas.

Si bien no existió una correlación entre los atributos extraídos y los resultados de F-score, se logró generar modelos que con un 70 % menos de atributos, obtienen resultados mayores a los del sistema base. Por otro lado, los resultados más bajos obtenidos siguen siendo competitivos, si se comparan con los expuestos en el WorkShop SemEval. Proponer este enfoque permitió además, prescindir de atributos que formaban parte esencial de los enfoques que se basan en los rasgos de superficie y lexicones como lo son los Chargrams. Siguiendo con esta idea, unos de los mejores resultados, obtenidos en esta investigación fue el resultante de la evaluación sobre el Corpus de test SemEval 2014, donde se obtuvo un 72,50 % en el F-score, con un modelo que prescindía de Chargrams y Ngrams, los que fueron reemplazados con los Ngrams de sentidos, en específico Nowsdgram, que son los obtenidos con el sentido más frecuente. Este modelo sólo utiliza un 21,24 % de los atributos en comparación a su experimento en la librería base.

Según lo expuesto por los autores del modelo original del cual se inspiró el modelo base, el prescindir de Chargrams y Ngrams degrada el sistema, hipótesis que para el enfoque propuesto no es válida. Utilizar un método de desambiguación en este caso, arroja resultados más altos a nivel global; sin embargo, utilizar el sentido más frecuente a nivel individual presenta los resultados más altos en esta investigación en 3 de los 4 Corpus evaluados.

Los resultados más bajos fueron los obtenidos por las librerías que prescindían de Chargrams, Ngrams y Clusters, dicha combinación es la menos beneficiada por este enfoque.

La efectividad del enfoque propuesto se evidenció a través de los resultados obtenidos en la etapa de experimentación, lo que llevó a seguir explorando el potencial de éste, compitiendo en la versión 2017 de SemEval. Como no se preparó un modelo en específico, se decidió crear un meta clasificador el cual funciona con 10 modelos con pesos de polaridad conocido, donde la polaridad final se otorga con un proceso de votación simple. Los modelos fueron seleccionados con el cálculo de un promedio por experimento en los años 2013 y 2014, generando un ranking, el cual se validó con el resultado de la Cross-validación de 30 de los primeros experimentos. El resultado de los 10 mejores modelos diferió en sólo un modelo entre ambos ranking por lo que se decidió evaluar el meta-clasificador sobre el corpus del año 2014 donde el mejor resultado lo logró el ranking generado por promedios.

Se consiguió el décimo lugar en la competencia con un 62.40 %; sin embargo, al probar los primeros 30 modelos de forma individual, se logran conseguir 29 modelos con un mejor F-score que el alcanzado por el meta-clasificador, donde destacan los modelos SB_SCG_SNG en el experimento 8, SB_SCG en el experimento 10 y SB_SCG en el experimento 11, que alcanzan respectivamente un 65,2 % y 65,4 para ambos últimos, lo que se traduce en un sexto lugar.

Estos lugares, dejan las puertas abiertas para preparar un modelo, que utilice este enfoque y se desempeñe en el área competitiva.

Referencias

Amir, S., Astudillo, R., Ling, W., Silva, M. J., and Trancoso, I. (2016). INESC-ID at SemEval-2016 Task 4-A: Reducing the Problem of Out-of-Embedding Words. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 238–242, San Diego, California. Association for Computational Linguistics.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 2200–2204.

Büchner, M. H. M. P. M. and Stein, B. (2015). Webis: An ensemble for twitter sentiment detection. *SemEval-2015*, 582.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly, Beijing ; Cambridge [Mass.], 1st ed edition.

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., De Luca, V., and Jaggi, M. (2016). SwissCheese at SemEval-2016 Task 4: Sentiment Classification Using an Ensemble of Convolutional Neural Networks with Distant Supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1124–1128, San Diego, California. Association for Computational Linguistics.

dos Santos, C. (2014). Think Positive: Towards Twitter Sentiment Analysis from Scratch. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 647–651, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Ebert, S., Vu, N. T., and Schütze, H. (2015). CIS-positive: A Combination of Convolutional Neural Networks and Support Vector Machines for Sentiment Analysis in

Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 527–532, Denver, Colorado. Association for Computational Linguistics.

Fellbaum, C. (2005). WordNet and wordnets. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford. Elsevier.

Fernández-Gavilanes, M., Álvarez López, T., Juncal-Martínez, J., Costa-Montenegro, E., and González-Castaño, F. J. (2015). GTI: An Unsupervised Approach for Sentiment Analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 533–538, Denver, Colorado. Association for Computational Linguistics.

Flekova, L., Ferschke, O., and Gurevych, I. (2014). UKPDIPF: Lexical Semantic Approach to Sentiment Polarity Prediction in Twitter Data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 704–710, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Gómez-Gómez, M., Danglot-Banck, C., and Vega-Franco, L. (2003). Sinopsis de pruebas estadísticas no paramétricas. cuándo usarlas. *Revista Mexicana de Pediatría*, 70(2):91–99.

Han, J. and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier, Burlington, MA, 3rd ed edition.

HLTCOE, J. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Atlanta, Georgia, USA*, 312.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., De Jong, F., and Kaymak, U. (2015). Exploiting Emoticons in Polarity Classification of Text. *J. Web Eng.*, 14(1&2):22–40.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Hunston, S. (2006). Corpus linguistics. *Linguistics*, 7(2):215–244.

Indurkhya, N. and Damerau, F. J. (2010). *Handbook of natural language processing*, volume 2. CRC Press.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Koppel, M. and Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Malandrakis, N., Falcone, M., Vaz, C., Bisogni, J. J., Potamianos, A., and Narayanan, S. (2014). SAIL: Sentiment Analysis using Semantic Similarity and Contrast Features.

In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 512–516, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill series in computer science. McGraw-Hill, New York, NY, international ed., [reprint.] edition.

Miura, Y., Sakaki, S., Hattori, K., and Ohkuma, T. (2014). TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

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

Negi, S. and Rosner, M. (2013). UoM: Using Explicit Semantic Analysis for Classifying Sentiments. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 535–538, Atlanta, Georgia, USA. Association for Computational Linguistics.

Nugues, P. M. (2006). *An introduction to language processing with Perl and Prolog: an outline of theories, implementation, and application with special consideration of English, French, and German.* Cognitive technologies. Springer, Berlin ; New York.

Ortega Bueno, R., Fonseca Bruzón, A., Gutiérrez, Y., and Montoyo, A. (2013). SSA-UO: Unsupervised Sentiment Analysis in Twitter. In *Second Joint Conference on*

*Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 501–507, Atlanta, Georgia, USA. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Potts (2011). Sentiment Symposium Tutorial: Tokenizing.

Proisl, T., Greiner, P., Evert, S., and Kabashi, B. (2013). KLUE: Simple and robust methods for polarity classification. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 395–401.

Reckman, H., Baird, C., Crawford, J., Crowell, R., Micciulla, L., Sethi, S., and Veress, F. (2013). teragram: Rule-based detection of sentiment phrases using SAS Sentiment Analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 513–519, Atlanta, Georgia, USA. Association for Computational Linguistics.

Rouvier, M. and Favre, B. (2016). SENSEI-LIF at SemEval-2016 Task 4: Polarity embedding fusion for robust sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 202–208, San Diego, California. Association for Computational Linguistics.

Ruder, S., Ghaffari, P., and Breslin, J. G. (2016). INSIGHT-1 at SemEval-2016 Task 4: Convolutional Neural Networks for Sentiment Classification and Quantification. In

Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 178–182, San Diego, California. Association for Computational Linguistics.

Severyn, A. and Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, pages 464–469.

Tang, D., Wei, F., Qin, B., Liu, T., and Zhou, M. (2014). Coooolll: A Deep Learning System for Twitter Sentiment Classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 208–212, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Wijksgatan, O. and Furrer, L. (2013). Gu-mlt-lt: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. *Atlanta, Georgia, USA*, 328.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human langua-*

ge technology and empirical methods in natural language processing, pages 347–354. Association for Computational Linguistics.

Xu, S., Liang, H., and Baldwin, T. (2016). UNIMELB at SemEval-2016 Tasks 4a and 4b: An Ensemble of Neural Networks and a Word2vec Based Model for Sentiment Classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 183–189, San Diego, California. Association for Computational Linguistics.

Zhang, Z., Wu, G., and Lan, M. (2015). ECNU: Multi-level Sentiment Analysis on Twitter Using Traditional Linguistic Features and Word Embedding Features. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 561–567, Denver, Colorado. Association for Computational Linguistics.

Zhu, X., Kiritchenko, S., and Mohammad, S. M. (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 443–447. Citeseer.