INSTITUTO POLITECNICO NACIONAL

CENTRO DE INVESTIGACION EN COMPUTACION

MEDIDAS DE ASOCIACIÓN EN ESCALAS BIPOLARES PARA LA MINERÍA DE OPINIÓN EN REDES SOCIALES

TESIS DOCTORAL

PRESENTA

M. en C. Fernando Monroy Tenorio

DIRECTOR DE TESIS

Dr. Alexander Gelbukh

Dr. Ildar Batyrshin

México, D.F. Junio de 2017

RESUMEN

Las opiniones son fundamentales para casi todas las actividades humanas y son factores de influencia en los comportamientos tanto de personas como de las organizaciones. Las creencias y percepciones de la realidad, y las decisiones que se toman, son, en gran medida, condicionadas de acuerdo a cómo los demás ven y valoran el mundo. Cuando se tiene que tomar una decisión muchas veces se contemplan las opiniones de los demás.

En Internet se encuentran grandes cantidades de información textual donde se vierten opiniones como: redes sociales, sitios de comercio electrónico (Amazon), o sitios especializados como tripadvisor en el cual un factor importante para que un usuario reserve un hotel son los comentarios. Podemos encontrar además una gran cantidad de información que refleja la opinión de los internautas recopilada a través de escalas de calificación en forma numérica o verbal estas se usan a menudo para medir las preferencias, rasgos, habilidades, actitudes y valores de los atributos humanos, pero también para calificar gustos por alguna película, un nuevo modelo de celular o algún libro, entre otros.

Generalmente, la escala de clasificación bipolar es un conjunto ordenado linealmente con la simetría entre los elementos considerados como categorías negativa y positiva. En este trabajo se discuten diferentes casos particulares de escalas bipolares. Se consideran sus estructuras típicas con etiquetas verbales que se utilizan para la construcción de este tipo de escalas. También se introduce y estudia el concepto de la función de puntuación bipolar preservando el ordenamiento lineal y la simetría de las escalas bipolares. Se proponen los métodos de construcción de las funciones de puntuación bipolares. Además se muestra que el coeficiente de correlación de Pearson a menudo utilizado para el análisis de las relaciones entre los perfiles de las calificaciones en los sistemas de recomendación puede ser engañoso si las escalas de calificación son bipolares.

En esta tesis se proponen nuevas medidas de correlación en escalas bipolares libres de los inconvenientes del coeficiente de correlación de Pearson. Estas medidas de correlación pueden utilizarse en el campo del análisis de sentimiento y la minería de opinión para el análisis de posibles relaciones entre las opiniones de los usuarios y sus calificaciones en productos o servicios.

También, se estudia el uso de nuevas características (features) para mejorar la clasificación de opiniones en español utilizando n-gramas sintácticos, evitando la perdida de información sobre las relaciones sintácticas que ocurre al utilizar n-gramas tradicionales como características en un enfoque de aprendizaje automático, los sn-gramas son extraídos de un corpus de críticas de películas en español y como algoritmo de aprendizaje se utilizan Maquinas de Soporte Vectorial para determinar la polaridad(positivo, negativo) de los textos, se compara la eficiencia del uso de los n-gramas sintácticos con n-gramas tradicionales para atacar esta problemática.

CONTENIDO TEMÁTICO

1. INTRODUCCIÓN 6

1.1. Justificación 7

1.2. Estructura del documento 8

2. ANTECEDENTES 10

2.1 Minería de opinión 10

2.2 Clasificación de polaridad 13

2.3 Selección de Características 17

2.4 Escalas de medición 19

3. MARCO TEORICO 22

3.1 Análisis sintáctico. 22

3.2 n-gramas. ..... **¡Error! Marcador no definido.**

3.2 n-gramas sintácticos. 23

3.3 técnicas de clasificación en aprendizaje automático. 23

3.4 Tipos de escalas bipolares 24

3.5 Escalas de calificación verbales bipolares 27

3.6 Escalas bipolares finitas 30

3.7 Funciones de puntuación bipolares 32

4. METODOLOGÍA 36

4.1 Selección del corpus 36

4.2 Análisis sintáctico 38

4.3 Algoritmo para la extracción de bigramas y trigramas sintacticos 39

4.4 Creación de matriz termino documento 43

4.5 Selección de clasificadores 45

4.6 Generación del conjunto de datos para entrenamiento y evaluación 46

4.7 Análisis sintáctico de las críticas del corpus 46

4.8 Métodos de construcción de las funciones de puntuación bipolar 48

4.8.1 Funciones de utilidad bipolares basadas en la distribución de las calificaciones de usuario 48

4.8.2 Funciones de utilidad basadas en generadores 51

4.8.2 Selección paso a paso de los valores del generador 53

5. METODOLOGIA DE EVALUACIÓN 53

5.1 Medidas de correlación en el conjunto de perfiles bipolares 54

6. Discusiones y Conclusiones 60

7. REFERENCIAS BIBLIOGRÁFICAS 62

1. INTRODUCCIÓN

La *minería de opinión* denominada también *análisis de sentimiento*, se centra en tratar automáticamente información con opinión, lo que permite, entre otras cosas, extraer la polaridad (positiva, negativa, neutra o mixta) de un texto (Pang y Lee, 2008).

La *minería de opinión* toma relevancia en nuestros tiempos debido a que la cantidad de información textual vertida en términos de opinión sobre productos y servicios es tan grande que hace imposible que los datos puedan analizarse manualmente. Al realizar una compra en línea la gente no llega a ver los productos reales y suele guiar sus preferencias por los comentarios que se han hecho por otros usuarios (Pan,2012).

Se han realizado múltiples trabajos relacionados con la minería de opinión sin embargo como menciona Liu (2010), monitorizar las opiniones de la Web es una tarea difícil y los modelos actuales del problema, así como la información necesaria o de sus posibles soluciones, son todavía limitados.

En la psicología, la sociología, la medicina y otros campos, las escalas de calificación en forma numérica o verbal se usa a menudo para medir las preferencias, rasgos, habilidades, actitudes y valores de los atributos humanos (Thurstone, 1928; Likert, 1932; Mosier, 1941; Osgood, 1952; Jupiter, 1999; Friborg, 2006). Pueden considerarse diferentes definiciones de la escala de calificación. En la actualidad, es más común considerar como una escala de calificación el conjunto de categorías ordenadas linealmente. El número de categorías utilizadas en las escalas de calificación por lo general varía de 3 a 11 (Likert, 1932; Mosier, 1941; Hjermstad, 2011). El conjunto de escalas de calificación puede ser agregado en la escala de nivel superior (Likert, 1932).

Las escalas de clasificación a menudo tienen una estructura bipolar: dos polos y categorías opuestas simétricamente situadas en lados opuestos de la escala (Thurstone,1928; Likert, 1932; Osgood,1952; Pospelov,1989; Petrenko,1988; Batyrshin,1990; Tarassov,2001; Dubois,2006). En tales escalas bipolares el lado negativo de la escala es el espejo inverso del positivo (Dubois, 2006). Un objeto se evalúa en escala bipolar como positivo o negativo o neutro.

Existen enfoques en los cuales se analiza la información textual vertida por el usuario para obtener al final una polaridad de su opinión como pang (2008), pero también existen trabajos en los cuales se evalúa la información contenida en escala de calificación como el caso de pang (2005).

En este trabajo se abordan ambos enfoques al estudiar las escalas de calificación de tipo bipolar y utilizar nuevas características (features) para la clasificación de opinión de textos en el aprendizaje máquina, se explora en este último enfoque el uso de n-gramas sintácticos (sn-gramas) propuestos por Sidorov(2013) en la clasificación de críticas de películas, se analizan los resultados obtenidos y se realiza una comparación con el uso tradicional de n-gramas como características en esta tarea.

1.1. *Justificación*

La importancia de clasificar adecuadamente las opiniones según Cambria (2013) puede ser crucial al momento de tomar decisiones o de elegir entre varias opciones. Cuando esas decisiones implican recursos valiosos (por ejemplo, el gasto de tiempo y dinero para comprar productos y servicios), la gente a menudo se basa en experiencias pasadas de otros. La web social ofrece nuevas herramientas para crear de manera eficiente y compartir ideas y opiniones con todo el mundo a través de internet.

La clasificación de polaridad se ha atacado desde el enfoque basado en técnicas de aprendizaje automático y enfoques basados en lexicones. La mayoría de los estudios se han centrado en el entrenamiento de clasificadores basados en algoritmos de aprendizaje automático para clasificar comentarios (Wang, 2014).

En gran parte de los trabajos donde se utiliza el enfoque basado den aprendizaje automático se utilizan palabras como características de entrenamiento con lo cual como menciona Sidorov (2013) se pierde información sobre las relaciones sintácticas entre palabras. Las palabras se convierten en lo que se llama *"bolsa de palabras"*. Otras características utilizadas para la clasificación de opiniones son los n-gramas tradicionales los cuales son secuencias de palabras tal como aparecen en un documento (Manning, 1999).

En este estudio se utilizan sn-gramas como características para evitar la pérdida de información sintáctica en un enfoque basado en aprendizaje automático y con esto mejorar la clasificación de la polaridad. Según Sidorov (2013). Los sn-gramas pueden ser utilizados como características de clasificación de la misma manera que n-gramas tradicionales, por lo tanto en cualquier tarea que se utilicen n-gramas tradicional se pueden aplicar también sn-gramas. La desventaja de los sn-gramas es que requieren del análisis sintáctico previo. La idea de los sn-gramas se justifica con la introducción de información lingüística a métodos estadísticos, con esto se pretende superar la desventaja tradicional de n-gramas de contener muchos elementos arbitrarios, es decir, una gran cantidad de ruido.

En muchas aplicaciones de las escalas de calificación, un ser humano es considerado como un indicador para medir y calificar la fuerza de las actitudes o valores de los atributos; Los resultados de tales "mediciones" son usualmente representados por números. La teoría de la medición (Pfanzagl,1971; Roberts,1985) estudia las clases de operaciones permitidas en los conjuntos de tales números, definiendo escalas ordinal, de intervalo y de medición de relación. Aunque las escalas de calificación suelen tener forma bipolar simétrica definida por un par de términos polares (Osgood,1952), esta simetría por lo general no se considera explícitamente en las escalas de medición. Debido al creciente interés en la aplicación de escalas de calificación bipolares en los sistemas de recomendación y en el análisis del sentimiento (Resnick, 1994; Adomavicius, 2005; Ricci, 2011; Liu, 2009; Tarassov, 2001; Cambria, 2014; Poria, 2014), el problema de considerar las funciones de puntuación definidas en las escalas bipolares teniendo en cuenta explícitamente la simetría de estas escalas, que estudiamos en este trabajo, es de particular interés.

1.2. *Estructura del documento*

El presente documento se organiza de la siguiente manera: en el *capítulo 1* se justifica la importancia de la investigación. En el *capítulo 2* se realiza una revisión del estado del arte relacionado a los enfoques de investigación por una parte el uso de nuevas características de clasificación en el aprendizaje automático para la tarea de minería de opinión y por otra el uso de medidas de asociación en perfiles de usuarios basados en puntuaciones de valoración a partir de escalas bipolares de clasificación.

En el *capítulo 3* se realiza la revisión de la literatura o estado del arte para establecer los antecedentes de trabajos relacionados con el tema de investigación.

El marco teórico descrito en el *capítulo 4* presenta información acerca del proceso de análisis sintáctico de dependencias, n-grama sintáctica y del clasificador SVM, también se aborda información referente .a las escalas de medición

En el *capítulo cinco* se muestran los pasos metodológicos que se realizaran para llevar a cabo la propuesta de investigación, en el capítulo cinco se menciona la metodología de evaluación propuesta, el *capítulo seis* aborda las conclusiones y discusiones y el último capitulo muestra las referencias bibliográficas consultadas para el desarrollo de la presente propuesta.

2. ANTECEDENTES

En este apartado se analiza el estado del arte del tema de investigación con el fin de establecer un panorama sobre los estudios relacionados con el problema de investigación.

2.1 Minería de opinión

Dentro del Procesamiento de Lenguaje Natural (PLN) la subjetividad es la expresión lingüística de cualquier opinión, sentimiento, emoción, evaluación, creencias y especulaciones (Wiebe, 1994).

Pang and Lee (2008) mencionan que los términos "*análisis de sentimientos*" y "*minería de opiniones*" hacen referencia a una misma tarea y que éstas son un subconjunto de una tarea mayor que es el *análisis de subjetividad*. Asimismo Wilson (2009) entiende la tarea del *análisis de sentimientos* como un subconjunto del *análisis de subjetividad*.

La *minería de opinión* también conocida como *análisis de sentimiento* es una tarea desafiante que combina técnicas de Minería de Datos y PLN para tratar computacionalmente la subjetividad en documentos de textos (Martín, 2012).

La *minería de opinión* se centra en tratar automáticamente información con opinión, lo que permite, entre otras cosas, extraer la polaridad (positiva, negativa, neutra o mixta) de un texto (Pang y Lee, 2008).

El termino *minería de opinión* fue definido por Dave (2003) como "dado un conjunto de documentos de texto evaluativos D que contienen la opinión acerca de un "objeto" (persona, organización, producto, etc.), la *minería de opinión* tiene como objetivo extraer los atributos y componentes del objeto que se han comentado en cada documento d en el conjunto D y determinar si los comentarios son positivos, negativos o neutrales".

En los párrafos anteriores Pang y Lee (2008), Martin (2012) mencionan que el *análisis de sentimiento* y la *minería de opinión* son la misma tarea , autores como Ortiz (2010), Martínez (2011), Vilares (2013) coinciden indicando que la *minería de opinión* es también conocida como *análisis de sentimiento*.

Sidorov (2013) indica que cuando se dice *opinión* se puede referir a un sentido mucho más amplio, apelando a las características sustanciales del objeto como por ejemplo el tamaño de un producto, su peso, etc. al decir *sentimiento* se refiere a sentimientos solo positivos y negativos, si se quiere analizar sentimientos más detallados se diría análisis o *minería de emociones*.

La relevancia de la *minería de opinión* se da por la expansión de Internet y la Web 2.0 como blogs, tweets, prensa en línea, páginas web de diferentes servicios hoteles, restaurantes, libros, etc. (Ortiz,2010). Dentro de estas páginas la diversidad y cantidad de críticas presentes resultan de gran utilidad a fabricantes y vendedores, que ven en ella un mecanismo para conocer de primera mano cómo sus artículos son percibidos por los consumidores (Vilares, 2013).

Muchas personas escriben sus opiniones en foros, microblogs o sitios web de opinión. Esta información es muy útil para las empresas de negocios, gobierno e individuos, que quieren seguir las actitudes y sentimientos en forma automática en estos sitios. Es decir, hay una gran cantidad de datos disponibles que contienen mucha información útil, por lo que puede ser analizada de forma automática (Sidorov, 2013).

Según Pang y Lee (2008) los primeros trabajos relacionados con la minería de opinión son los de Carbonell (1979) y Wilks (1983) en el primero se habla de la *subjetividad*, el autor indica que no hay dos personas que piensen exactamente igual, diferentes personas se suscriben a diferentes creencias y están motivadas por diferentes objetivos en sus actividades; en el segundo trabajo se habla acerca de *agrupamiento de creencias* en cuanto al mecanismo de entorno computacional.

El año 2001 parece marcar el comienzo de una amplia conciencia de los problemas de investigación y las oportunidades del *análisis de sentimiento* y la *minería de opinión*, posteriormente, ha habido literalmente cientos de artículos publicados sobre el tema (Pang y Lee, 2008).

Los factores detrás de esta “fiebre” según Pang y Lee (2008) son:

- El aumento de los métodos de Aprendizaje Maquina en el Procesamiento de Lenguaje Natural y la Recuperación de Información.
- La disponibilidad de datos de entrenamiento para los algoritmos de Aprendizaje Automático debido al florecimiento de la Word Wide Web y, en concreto, el desarrollo de sitios web de agregación de opinión.
- Realización de los desafíos intelectuales y aplicación comerciales y de inteligencia que ofrece el tema.

El análisis de sentimiento se ha aplicado en sistemas en línea para la detección de páginas web con contenido inapropiado como el realizado por Jin (2007). Otra área donde el análisis de sentimiento puede ser útil es el análisis de las respuestas de preguntas de opinión como el propuesto por Lita (2005). Dave (2003) analiza opiniones en productos electrónicos y de consumo, Das y Chen (2001) aplica el análisis de sentimiento a mercados financieros.

La mayoría de los trabajos de investigación relacionados con la minería de opiniones se centra en textos en el idioma inglés (Martínez, 2011). El aumento de información sobre opiniones en idiomas distintos al inglés ha motivado trabajos de investigación en el campo de la *minería de opinión* en idiomas como el español, chino, francés, alemán y otros.

Cruz (2008) realiza un corpus de opiniones de películas en idioma español, Zhang (2009) realiza su estudio en corpus del idioma chino, Ghorbel y Jacot (2010) enfocan su estudio a comentarios de películas en francés, Agic (2010) presentan un corpus de opiniones de mercados financieros en el idioma croata.

Existen también investigaciones que tratan de analizar textos en varios idiomas como Ahmad (2006) que trabaja sobre un corpus de noticias financieras en los idiomas inglés, chino y árabe. Boldrini (2009) presenta el corpus EmotiBlog el cual contiene comentarios en los idiomas español, inglés e italiano.

Otros estudios realizan traducción de las opiniones al idioma ingles para aprovechar los métodos y recursos existentes en este idioma y que no se encuentran en el idioma nativo de la opinión. Martín-Valdivia (2012) traduce el corpus realizado por Cruz (2008) de

español al idioma ingles para aprovechar sentiWordNet que es un recurso léxico para minería de opinión en inglés. Casos similares son los trabajos realizados por Boldrini (2009), Wan (2009).

En los últimos años los trabajos realizados en esta materia ha aumentado. Los artículos relacionados con el tema que se pueden encontrar son muy grandes. Como indica Strapparava y Milhacea (2007), la detección automática de la emoción en textos es cada vez más importante desde el punto de vista aplicativo.

Dentro de sus estudios Strapparava y Mihalcea (2007) han realizado minería de opinión sobre titulares de noticias en la Web y análisis de sentimiento multimodal, en este último se integran características visuales, de audio y de texto que son utilizadas para una detección efectiva de sentimientos en video.

2.2 *Clasificación de polaridad*

El término "polaridad" consiste en determinar si una palabra comunica una actitud positiva o negativa, se encuentra en verbos, pronombres, adjetivos y adverbios. Puede cambiar la connotación de una frase o un texto (Polany y Zaenen, 2004).

La *minería de opinión* incluye una gran cantidad de tareas que han sido tratadas en mayor o menor medida (Pang y Lee, 2008). Una de las principales aplicaciones consiste en determinar la polaridad de las opiniones a nivel de documento, frase o característica. De esta manera podemos clasificar binariamente las opiniones en positivas o negativas (Martínez, 2011).

Un número considerable de documentos que citan el término "análisis de sentimientos" se centran en la aplicación específica de clasificar críticas en cuanto a su *polaridad*, un hecho que parece haber causado a algunos autores sugerir que la frase se refiere específicamente a esta tarea. Sin embargo muchos interpretan el termino para referirse al tratamiento computacional de la opinión, el sentimiento y la subjetividad del texto (Pang y Lee, 2008).

Autores como Martin-Valdivia (2012), Wang (2010) coinciden en que existen dos enfoques o metodologías que se han utilizado para la clasificación de la *polaridad*. El

primero es el enfoque de *aprendizaje automático* (supervisado) que se basa en el uso de una colección de datos para el entrenamiento de un clasificador. El segundo enfoque *orientación semántica* (no supervisado) no requiere de un entrenamiento previo sino que se tiene en cuenta la orientación positiva o negativa de las palabras.

Dentro del enfoque de *orientación semántica* uno de los métodos más mencionados en la literatura es el propuesto por Turney (2002) donde utiliza un algoritmo de aprendizaje no supervisado para determinar la orientación semántica de los términos de manera individual.

En el trabajo de Turney (2002) la *orientación semántica* de una frase se calcula como la información mutua entre la frase dada y la palabra. Se centra en el análisis de adjetivos y adverbios los cuales son extraídos del texto de entrada usando la propuesta de Brill (1994).Una vez identificadas las frases que tienen adjetivos y adverbios se estima la *orientación semántica* de cada frase extraída.

Para extraer la orientación semántica de cada frase Turney (2002) utiliza el algoritmo PMI (*pointwise mutual information*). Este algoritmo utiliza la información mutua como una medida de la fuerza de asociación semántica entre dos palabras (Church, 1990). El puntaje PMI de dos palabras p_1 y p_2 esta dada por la probabilidad de que dos palabras estén ocurran juntas dividida entre la probabilidad de ocurrencia de cada palabra y se expresa con la siguiente ecuación:

$$PMI_{(p1,p2)} = \log \frac{P(p_1,p_2)}{P(p_1)P(p_2)} \quad (1)$$

La fórmula para la orientación semántica de una palabra se obtiene con la siguiente ecuación:

$$SO - PMI_{(palabra)} = PMI(palabra, p_query) - PMI(palabra, n_query)$$

(2)

Donde los términos de referencia positivos y negativos son

$$p_query = \ bueno\ OR\ bonito\ OR\ exelente\ OR\ positivo\ OR\ afortunado\ OR\ correcto\ OR\ superior$$

$$n_query = malo\ OR\ repugnante\ OR\ pobre\ OR\ negativo\ OR\ desafortunado\ OR\ mal\ OR\ inferior$$

OR y NEAR son operadores ofrecidos por AltaVista search enginem. Por aproximación de PMI con el número de aciertos obtenidos por el motor de búsqueda y haciendo caso omiso al número de documentos del corpus (Kennedy, 2006), la formula se transforma:

$$SO - PMI = \log \frac{hits(palabra\ NEAR\ p_query)\ hits(n_query)}{hits(palabra\ NEAR\ n_query)\ hits(p_query)} \quad (3)$$

La *orientación semántica* de términos y frases se utiliza para determinar el sentimiento de sentencias completas y comentarios. Se calcula el promedio de orientación semántica de las frases y términos y se clasifica el comentario en positivo y negativo (Turney, 2010).

Al método propuesto por Turney (2012) se le conoce también como el método de conteo de termino, el cual determina que si existen más términos positivos que negativos el comentario es positivo, si existen más términos negativos que positivos el comentario es negativo, si existe el mismo número de términos positivos y negativos el comentario es neutro (Kennedy, 2006).

Este método también conocido como el enfoque no supervisado, hace uso de recursos léxicos como listas de palabras etiquetadas con su polaridad, lexicones, diccionarios, etc. la orientación semántica requiere una gran cantidad de recursos lingüísticos que por lo general dependen de la lengua y con frecuencia este enfoque obtiene menor recall, ya que depende de la presencia de las palabras que comprenden el léxico en el documento con el fin de determinar la orientación en la opinión (Matín-Valdivia, 2012).

Dentro del enfoque de *orientación semántica* Kim y Hovy (2004), formulan modelos para la clasificación de la polaridad a nivel de oración en función de las polaridades e intensidades de las expresiones polares de cada oración.

Moreno (2012), propone el sistema de análisis de sentimientos *Sentitext*. El sistema contiene una base de datos de palabras con más de 10,000 entradas, a cada palabra se le asigna una valencia –2, –1, 1, o 2, dependiendo si su carga es muy negativa, negativa, positiva o muy positiva. La inserción de palabras se realizó utilizando el diccionario de

sinónimos en español de OpenOffice. Las valencias se asignan en base a la semántica de la palabra y con la ayuda de un corpus de texto de lenguaje general.

El enfoque de *aprendizaje automático* (supervisado) se basa en el uso de una colección de datos en el caso de la clasificación de la polaridad corpus con reseñas o comentarios etiquetados con su polaridad para el entrenamiento de clasificadores (Matín-Valdivia, 2012). Los datos de entrenamiento y reales se procesan antes de ser utilizados por el clasificador con el fin de corregir errores y extraer las características principales (Fernandez, 2012).

Pang (2002) realizo experimentos con algoritmos de aprendizaje supervisados como son: *Naive Bayes*, *Maximun Entropy* y *Suppor Vector Machine* (SVM). Utilizando como atributos de los vectores de aprendizaje las frecuencias de los términos, la presencia o no de estos, el uso de adjetivos, n-gramas y las categorías gramaticales de los términos.

Este enfoque depende de la disponibilidad de conjuntos de datos etiquetados para el entrenamiento, que en muchos casos son difícil de conseguir, en parte debido a la novedad de la tarea (Matín-Valdivia, 2012).Trabajos como el de Pang (2002) reportan mejores resultados utilizando el clasificador SVM al igual que Go (2009).

La dificultad principal de los métodos supervisados se encuentra en como representar los datos de entrada para generar el vector de cada instancia. Existen otras formas de representar los valores de entrada del vector de aprendizaje como son en términos, dependencias sintácticas y negación.

El uso de palabras como variables en los vectores de aprendizaje ha sido la forma más utilizada para representar los datos de entrada en el clasificador, se ha utilizado su frecuencia, el uso de n-gramas, de lemas o incluso su posición en el texto ejemplo de esto son los trabajos realizados por Pang (2002), Dave (2003), Mullen (2004) y Aue (2005).

Para establecer las frecuencias de las palabras se ha utilizado tf e idf como es el caso de Martínez (2011) y Martineau (2009). También se han utilizado dependencias sintácticas en los vectores de clasificación como el caso de Wilson (2005) y Abbasi (2008).

Existen trabajos que combinan ambos enfoques como el caso del trabajo propuesto por Matín-Valdivia (2012) donde ocupa un meta clasificador para combinar ambos enfoques dentro del enfoque supervisado utiliza SVM y como entrada los vectores de características son y como método supervisado el conteo de términos propuesto por Turney (2002). El algoritmo utilizado como meta clasificador es el denominado sistema de votos propuesto por Jhonson (2005).

2.3 *Selección de Características*

Dentro de los trabajos relacionados con el problema de investigación se han encontrado muy pocos que trabajan clasificación de polaridad en el idioma español. De estos se han analizado aquellos que tienen un enfoque basado en aprendizaje automático para estudiar qué tipos de características utilizan para el entrenamiento de sus clasificadores.

Martinez-Camara (2011), utilizan un enfoque de aprendizaje automático aplicado a corpus en español MuchoCine (Cruz 2008), utilizan como clasificadores Maquinas de Soporte Vectorial y Naive Bayes, como características utilizan el enfoque de bolsa de palabras y realizan variaciones de está quitando palabras vacías (Stop Words) y obteniendo la raíz de las palabras para reducir su espacio de dimensiones. Obteniendo mejores

Resultados al utilizar la bolsa de palabras sin las palabras vacías y como pesos la medida TF-IDF, en ambos clasificadores. Este experimento se extiende en varios experimentos como es el caso de Martínez-Valdivia (2012), Malvar-Fernández (2011) utilizando el mismo tipo de características.

Perea-Ortega (2014), realiza un estudio en el cual utiliza diversas variantes en cuanto a las características aplicado a la clasificación de tuits en español. Experimenta con diferentes combinaciones de características, tales como unigramas, bigramas, skipgramas(no considerando palabras vacías ni números) y mezclas entre ellos. Los mejores resultados que obtiene es remplazando los emoticos con bigramas.

Fernandez (2012), realiza un estudio para la clasificación de polaridad y detección de tema utilizando un enfoque de aprendizaje automático, como características utiliza n-gramas en un enfoque de bolsa de palabras.

Sidorov (2013), realiza un estudio para la clasificación de tuits en español basado en un enfoque de aprendizaje automático donde utiliza como características n-gramas donde n=1 hasta n=6, para diferentes clasificadores obteniendo mejores resultados utilizando Maquinas de Soporte Vectorial y unigramas.

En idiomas como el inglés se ha utilizado también con bastante frecuencia el uso de n-gramas como características tal es el caso de Pang (2002) en el cual comparan diferentes técnicas de aprendizaje automático y diferentes tipos de características como son unigramas, bigramas , unigramas mas bigramas; reportando mejores resultados el uso de unigramas y Maquinas de Soporte vectorial en la clasificación de polaridad de críticas de películas en inglés.

Contrario a lo reportado pór Pang (2002), Dave(2003) obtiene mejores resultados utilizando bigramas y trigramas en la clasificación de polaridad de críticas de productos.

Otro trabajo en el cual se usan varios tipos de características para entrenamiento es el de pak (2010) en el cual se utilizan unigramas, bigramas y trigramas para la clasificación de la polaridad en tuits en inglés. Encontrando mejores resultados en el uso de bigramas con una configuración especifica incorporando el uso de negación.

Recientemente Le (2015) utilizan unigramas, bigramas y un set de características denominado características orientadas a objetos (object-oriented features). Dentro del uso de bigramas utilizados en este trabajo se trata de eliminar de los bigramas aquellas palabras que no tienen un sentido o que no aportan información de sentimiento al clasificador.

Como se muestra en la revisión de los trabajos relacionados con la tarea de clasificación de polaridad el uso de n-gramas son comúnmente utilizados como características dentro del aprendizaje automático. Sin embargo como menciona Grigory(2013) los n-gramas

2.4 *Escalas de medición*

En muchas aplicaciones donde se usan escalas de calificación, como el análisis de sentimiento, el ser humano es considerado como un indicador para medir y calificar la fuerza de las actitudes o valores de los atributos y los resultados de tales "mediciones" son usualmente representados por números. La teoría de la medición (Pfanzagl, 1971; Roberts, 1985) estudió las clases de operaciones permitidas en los conjuntos de tales números que definen ordinal, intervalo y relación de las escalas de medición. Aunque las escalas de calificación por lo general tienen una forma bipolar simétrica definida por un par de términos polares (Osgood, 1952), esta simetría por lo general no se considera explícitamente en las escalas de medición. Debido al creciente interés en la aplicación de escalas de calificación bipolares en los sistemas de recomendación y en el análisis del sentimiento (Resnick, 1994; Adomavicius, 2005; Ricci, 2011; Liu, 2009; Taboada, 2011; Cambria, 2014; Poria, 2014), el problema de considerar las funciones de puntuación definidas en las escalas bipolares teniendo explícitamente en cuenta la simetría de estas escalas son de particular interés y estudiadas en este trabajo.

Alternativamente al enfoque basado en la medición para el análisis de las actitudes humanas, considerado en l los modelos de palabras, las evaluaciones verbales humanas de los atributos, las puntuaciones del sentimiento y las calificaciones. Por ejemplo, en la lógica difusa y la computación blanda (Zadeh, 1975; Jang, 1997), las palabras que representan la fuerza de los atributos como pequeño y muy bueno son modeladas por conjuntos difusos y muchas veces definidas por funciones paramétricas: triangulares, trapezoidales, etc. De la toma de decisiones, la clasificación o el control de los términos lingüísticos representados por conjuntos difusos se suelen considerar como los componentes de los sistemas de inferencia difusa y estos conjuntos difusos se puede ajustar mediante algún procedimiento de aprendizaje automático para obtener las soluciones óptimas o razonables en la salida de estos sistemas (Jang, 1997).

Herrera (2000) considera la estructura ordenada de términos lingüísticos con el operador de negación definido en el conjunto de índices. Esta operación de negación define la bipolaridad de la escala. La semántica del término lingüístico establecido en (Herrera, 2000) está dada por conjuntos difusos definidos en [0,1] que están distribuidos

simétricamente o no simétricamente en [0,1]. Xu (2012) considera las denominadas escalas de evaluación lingüística aditiva donde las puntuaciones son dadas por índices fraccionales de etiquetas lingüísticas y la bipolaridad de la escala es presentada por la negación también definida en el conjunto de índices.

En los sistemas de recomendación y el filtrado colaborativo la función de puntuación puede considerarse como función de utilidad típicamente representada por las calificaciones (Adomavicius, 2005). Las escalas de calificación suelen tener forma bipolar, pero en general, la utilidad puede ser una función arbitraria. Las extrapolaciones de las clasificaciones conocidas a desconocidas en los sistemas recomendadores se realizan generalmente (a) especificando heurísticas que definen la función de utilidad y validan empíricamente su desempeño, y (b) estimando la función de utilidad que optimiza cierto criterio de desempeño (Breese,1998) Consideremos dos modelos probabilísticos alternativos para el filtrado colaborativo, el clasificador bayesiano y las redes bayesianas. Los algoritmos de aprendizaje de máquina se aplican para aprender parámetros de modelos.

Liu (2009) considera el problema de asignar una puntuación para el grado de sentimiento polar a la frase relacionada con los aspectos considerados de los elementos. Basándose en estas puntuaciones, se calcula una calificación promedio para el aspecto. Los autores propusieron un modelo aditivo lineal para calcular las puntuaciones de sentimientos polares tomando valores en escala positiva. Taboada(2011) presentan un enfoque basado en el léxico para extraer el sentimiento del texto. Se utilizan los diccionarios de palabras anotadas con su orientación semántica (polaridad y fuerza). La orientación semántica (SO) de las palabras toma valores positivos y negativos. Se supone que algunos problemas de cálculo de SO probablemente podría resolverse ajustando los valores de orientación sentimental y modificadores (Taboada, 2011).

Varios recursos en línea se desarrollan los últimos años para el análisis del sentimiento de los textos. Por ejemplo, SenticNet (2014) proporciona puntuaciones de polaridad para 30,000 conceptos Poria(2014). La puntuación oscila entre -1 (mala) a +1 (buena), con puntuaciones neutrales alrededor de cero.

Para analizar la similitud entre las listas de calificaciones en los sistemas de recomendación se utiliza a menudo el coeficiente de correlación de Pearson (Ricci, 2011). Sin embargo, se puede demostrar que este coeficiente de correlación puede ser engañoso en el análisis de las calificaciones de las escalas bipolares. Shardanand y Maes (1955) propusieron utilizar el coeficiente de Pearson "limitado" r para tener en cuenta la bipolaridad de la escala de calificación.

3. MARCO TEORICO

En este capítulo se presenta información conceptual relacionada con la propuesta de investigación con la finalidad de interpretar los pasos metodológicos que se describen más adelante.

3.1 *Análisis sintáctico.*

La tarea principal del análisis sintáctico es de describir como las palabras de la oración se relacionan y cuál es la función que cada palabra realiza en esta oración, es decir, construir la estructura de la oración de un lenguaje (Haro, 2007).

Para el caso de la propuesta las relaciones importantes en este nivel son la de los términos con sentimiento y los CVS, como en el caso de la frase "ninguno de ellos tiene éxito" la palabra ninguno es un tipo de negación y el termino de sentimiento en la oración es éxito el cual se puede clasificar con una polaridad positiva.

Dentro del análisis sintáctico existen dos enfoques el primero es el análisis por constituyentes y el segundo el análisis de dependencia. En el primer enfoque la estructura de la oración se describe agrupando palabras y especificando el tipo de cada grupo, usualmente de acuerdo con su palabra principal, estos grupos pueden ser por ejemplo: grupo verbal, grupo nominal (Chomsky, 2002).

En enfoque de dependencias, las dependencias se establecen entre pares de palabras, donde una es principal o rectora y la otra está dependiente de la primera. Si cada palabra de la oración tiene una propia palabra rectora, la oración entera se ve como una estructura jerárquica de diferentes niveles, como un árbol de dependencias. La única palabra que no está subordinada a otra es la raíz del árbol (Haro, 2007).

En la propuesta de investigación se utiliza el enfoque de dependencias, para saber la relación de un lexema con los distintos grupos de palabras en la oración, la descripción de los medios linguisticos con los que se expresan los objetos del lexema se insertan con el junto con un diccionario (Haro, 2007).

3.2 *n-gramas sintácticos.*

Los n-gramas para el PLN son una secuencia de palabras o elementos contenidos en el texto, la n corresponde al número de elementos en la secuencia, de este modo se tienen unigramas, bigramas, trigramas y n-gramas.

La diferencia entre los n-gramas y los sn-gramas (n-gramas sintácticos) es la forma en que elementos son considerados como vecinos. En el caso de los sn-gramas, los vecinos se toman siguiendo relaciones sintácticas en los arboles sintácticos, y no mediante la adopción de las palabras tal y como aparecen en el texto (Sidorov, 2013).

La ventaja de los sn-gramas es que cada palabra se une a su vecino real ya que se basa en las relaciones sintácticas como se mencionó anteriormente, haciendo caso omiso de la arbitrariedad que se introduce por la estructura de la superficie (Sidorov, 2013).

Según Sidorov (2013). Los sn-gramas pueden ser utilizados como características de clasificación de la misma manera que n-gramas tradicionales, por lo tanto en cualquier tarea que se utilicen n-grammas tradicionales se pueden aplicar también ns-grammas. La desventaja de los sn-gramas es que requieren del análisis sintáctico previo. La idea de los sn-gramas se justifica con la introducción de información lingüística a métodos estadísticos, con esto se pretende superar la desventaja tradicional de n-gramas de contener muchos elementos arbitrarios, es decir, una gran cantidad de ruido.

3.3 *técnicas de clasificación en aprendizaje automático.*

Para el PNL y en específico para la tarea de clasificación de la polaridad se utiliza en el enfoque orientado al aprendizaje automático clasificadores entre los que se encuentran Nayves Bayes (NV), Suport Vector Machine (SVM), maximun entropy , entre otros.

La técnica de clasificación de aprendizaje automático Nayves Bayes propuesta por Mitchell (1997), se fundamenta en la teoría de Bayes. Existen diversos métodos de estimación como el estimador de densidad de Kernel el cual varía su comportamiento en función al parámetro del ancho de banda.

El algoritmo de clasificación SVM separa en dos clases un conjunto de datos y ue propuesto por Vapnik (1998). Se aplica a problemas de clasificación en una infinidad de dominios, cuenta con diversas opciones de configuración, algunas sus ventajas son:

Eficiencia en los espacios dimensionales elevados.

- Eficaz en los casos en que el número de dimensiones es mayor al número de muestras
- Utiliza un subconjunto de puntos de entrenamiento en la función de decisión llamada vectores de soporte

Estos dos clasificadores han sido utilizados en la tarea de clasificación de la polaridad, Pang y Lee (2012) en un comparativo entre clasificadores reporta que Nayve Bayes y SVM tienen un mayor grado de precisión en la realización de esta tarea.

En la actualidad es fácil implementar en aplicaciones de PLN estos clasificadores un ejemplo es WEKA que es una herramienta que contiene una colección de algoritmos de aprendizaje automático, el cual es de código abierto. La librería de NLTK de pahyton contiene un paquete de nombre Classify en el cual se encuentran estos dos clasificadores ya programados.

3.4 Tipos de escalas bipolares

Generalmente, una escala bipolar es un conjunto ordenado linealmente con elementos mínimos y máximos considerados como dos polos de la escala. Estos polos pueden ser interpretados como polos negativo y positivo correspondientemente. Por lo general, las escalas bipolares tienen una categoría neutra situada en el centro de la escala y todas las demás categorías entre la categoría neutra y los polos negativo y positivo pueden considerarse como categorías negativa y positiva correspondientemente.

Además del ordenamiento lineal de las escalas bipolares también se da la simetría de la escala de tal manera que las categorías negativas se asignan a las categorías positivas opuestas y viceversa. Los elementos de las escalas bipolares pueden tener puntuaciones numéricas y etiquetas lingüísticas. La simetría de las escalas bipolares puede reflejarse en la simetría de las etiquetas y/o en la simetría de las puntuaciones. Los ejemplos típicos de

escalas bipolares que contienen tres o cinco respuestas posibles en las preguntas se consideran en (Likert,1932).

Una tipo de escalas con estructura bipolar por ejemplo puede tener cinco categorías: fuertemente aprobado, aprobado, indeciso, desaprobado, fuertemente rechazado, con puntuaciones 5, 4, 3, 2, 1 o invertir el orden. Esta escala tiene categoría neutral indecisa y simetría entre categorías positivas y negativas. Actualmente es más popular ordenar las categorías de la escala de negativo a positivo, por ejemplo, como sigue: fuertemente rechazado, desaprobado, ni en acuerdo ni en desacuerdo, de acuerdo, totalmente de acuerdo, o en correspondencia con el orden de las puntuaciones: 1, 2, 3, 4, 5.

Diferentes tipos de las escalas bipolares se consideran a continuación y se proporcionan por algunas referencias que contienen los ejemplos correspondientes:

A. Con número finito n de categorías (Likert, 1932).
B. Con un número infinito de categorías, (por ejemplo, [0,1], [-1,1], [-10,10]) (Herrera,2000; Grabisch,2006; Goldberg,2001).
C. Escala de 2 puntos (n = 2) (Schafer, 2007) o escala multipunto (n> 2).
D. Con categoría neutra (Likert, 1932) o sin categoría neutra (Chang, 1994).
E. Con etiquetas verbales sólo en los polos (Osgood, 1952) o totalmente etiquetados cuando todas las categorías de respuesta están explícitamente etiquetados (Likert, 1932; Weijters, 2010).
F. Con etiquetas simbólicas de polos (Petrenko, 1988).
G. Con puntuaciones numéricas expresamente dadas (Likert, 1932) o no, por ejemplo dado gráficamente con intervalos (Osgood, 1952).
H. Con puntuaciones numéricas positivas (por ejemplo 1, ..., 5, o [0,1]) (Likert, 1932; Herrera,2000) o con puntuaciones tanto negativas como positivas (por ejemplo, [-1,1] (Grabisch,2006; Cambria,2014).
I. Sin polaridad de las etiquetas verbales, es decir, una de las escalas de calificación de Likert (1932) contiene respuestas: Escuela primaria, secundaria, secundaria, colegio, graduado y profesional con las calificaciones 1, 2, 3, 4, 5 típicas Para las escalas bipolares.

J. Con polaridad de categorías verbales pero sin simetría con respecto a la categoría neutra, es decir, cuando alguna categoría o concepto verbal positivo no se corresponde con la categoría verbal negativa opuesta (Cambria, 2014; Poria, 2014).

K. Explícitamente utilizando la operación de negación (Zadeh, 1975; Herrera, 2000; Xu, 2012) o no (Likert, 1932).

L. Con puntuaciones no numéricas, p. Con conjuntos difusos (Zadeh, 1975; Herrera, 2000).

M. Con categorías verbales ordenadas de categorías negativas a la izquierda a categorías positivas a la derecha o inversa en orden (Likert, 1932).

N. Con las puntuaciones numéricas aumentaron de izquierda a derecha 1, ..., 5 o revés en orden 5, ..., 1 (Likert, 1932).

O. Con función de puntuación lineal (Likert, 1932) o no lineal con respecto a los índices de las categorías (Taboada, 2011; Xu, 2012).

P. Con varias etiquetas verbales o conceptos asignados a una gradación de la escala (debido a sinónimos y / o tener el valor de puntuación igual) (Cambria, 2014; Ricci, 2011).

Q. Con categorías positivas y negativas separadas en dos escalas (Dubois,2006; Thelwall,2010).

R. La actitud o atributo puede tener una positiva y una polaridad negativa grados al mismo tiempo (por ejemplo, grado de pertenencia y grado de no pertenencia) (Atanassov,2005; Dubois,2006).

S. Con agregación de puntuaciones de la misma escala (Thurstone,1928; Grabisch,2006; Batyrshin, 2011).

T. Con la agregación de las puntuaciones de diferentes escalas (Likert, 1932, Liu,2009).

U. Con el cálculo de la opinión del grupo como una agregación de las puntuaciones de los individuos (Thurstone, 1928).

V. Con el cálculo de puntajes individuales en escala de nivel superior como agregación de puntuaciones individuales en escalas particulares (Likert, 1932).

W. Midiendo (Likert, 1932) o modelando (Herrera, 2000; Adomavicius, 2005) actitudes, preferencias, clasificaciones o utilidades.

En este trabajo se, consideran escalas bipolares finitas con elemento neutro como conjuntos ordenados linealmente con simetría dada por la operación de negación. En la siguiente sección, se discuten las estructuras típicas de las escalas bipolares finitas con etiqueta completa con etiquetas verbales. Además, consideramos formalmente dos tipos de escalas bipolares finitas mutuamente relacionadas con el elemento neutro como los conjuntos de índices enteros de categorías de escalas bipolares. Introducimos funciones de puntuación bipolares definidas en estos conjuntos como modelos de preferencias de usuario.

3.5 Escalas de calificación verbales bipolares

Una escala de calificación verbal es bipolar si es simétrica con respecto a las categorías opuestas situadas en los lados opuestos de la escala. Tal simetría puede expresarse mediante una operación de negación N definida en la escala. Considere un ejemplo de la escala bipolar de 5 puntos con las etiquetas verbales ordenadas de izquierda a derecha: *nunca <raramente <a veces <a menudo <siempre*, y con la negación: *N (nunca) = siempre, N (rara vez) N (a veces) = a veces, N (a menudo) = rara vez, N (siempre) = nunca.*

Esta operación de negación tiene varias propiedades formales. Involutiva, es decir, la doble negación de cualquier categoría da la misma categoría, por ejemplo. *N (N (rara vez)) = rara vez*. La operación de negación va disminuyendo, es decir, si una categoría c_2 tiene la mayor calificación que una categoría c_1: $c_1 < c_2$, entonces la negación de c_1 tiene la mayor calificación que la negación de c_2: $N(c_1) > N(c_2)$. Por ejemplo, *nunca <implica a menudo N (nunca) = siempre> N (a menudo) = raramente*. Tenga en cuenta que esta operación de negación difiere de la lingüística *no* (Liu, 2009; Taboada, 2011).

Si la escala tiene un número impar de categorías entonces hay un *centro, punto medio o categoría C neutro en la escala tal que* $N(C) = C$. Generalmente, tal punto se llama un *punto fijo de la negación*. Para la escala considerada arriba tenemos C = *a veces*.

Consideremos la estructura típica de las escalas bipolares verbales de 7 puntos (Batyrshin, 1990):

L= (eap, vap, ap, np, p, vp, ep)…… (1)

L es ordenado de izquierda a derecha *eap* <*vap* <... <*vp* <*ep*, donde *p* y *ap* (anti-p) denotan adjetivos o atributos opuestos, correspondientemente, y otras letras denotan: *e = extra, v = very, n = neutral*. La negación define la correlación entre las categorías opuestas, por ejemplo: *N (eap) = ep, N (vap) = vp, ..., N (ep) = eap*, con *N (np) = np* para la categoría neutra np. He aquí un ejemplo de escala verbal bipolar con la estructura (1):

L = (absolutamente horrible, muy malo, malo, no bueno, no malo, bueno, muy bueno, excelente),

Con *p = bueno, ap = malo*. Se puede considerar la estructura (1) como un armario con estantes con las etiquetas *eap, vap, ..., ep* ordenados desde abajo hacia arriba. En particular, cada estante de (1) se puede llenar con etiquetas verbales correspondientes. Estas estanterías pueden tener puntuaciones 1, 2, 3, 4, 5, 6, 7 o -3, -2, -1, 0, 1, 2, 3 con la categoría neutra *np* que tiene la puntuación 4 o 0, correspondientemente. En las secciones siguientes, consideraremos estas puntuaciones como índices de las categorías de la escala bipolar con *n> 3* índices y la función de puntuación será dada como una función numérica definida en el conjunto de estos índices. Para la función de puntuación de identidad sus valores coincidirán con los índices, en nuestro caso con 1, ..., 7 o con -3, ..., 3. Generalmente, será función no lineal preservando la simetría de la escala bipolar.

Los estantes de (1) pueden contener varias etiquetas verbales que se consideran como sinónimos o que tienen puntuaciones iguales. Por ejemplo, el estante *eap* puede contener las etiquetas verbales *absolutamente horrible* y *completamente terrible*. La escala bipolar de 5 puntos se puede obtener de la estructura (1) eliminando categorías simétricas o por fusión simétrica de categorías vecinas. De este modo, se puede reducir la estructura (1), por ejemplo, a la siguiente:

L= (vap, ap, np, p, vp),……..(2)

Con los índices 1, 2, 3, 4, 5 o -2, -1, 0, 1, 2. En el armario reducido (2) el estante *vp* puede contener las etiquetas verbales *muy bueno, excelente, perfecto etc*. La categoría *np* dará escala verbal bipolar sin centro.

Otra estructura típica de la escala bipolar tiene la siguiente estructura [3]:

L= (hap, map, lap, np, lp, mp, hp), ……..(3)

Donde *h, m* y *l* denotan intensidades correspondientemente *altas, medias* y *bajas* de los atributos *op* y *ap*. Las categorías de la escala (3) se ordenan de izquierda a derecha: *hap <map <... <mp <hp*. La negación define el mapeo entre las categorías opuestas: *N (hap) = hp, N (mapa) = mp, ..., N (hp) = hap*. Las escalas bipolares con el número de categorías inferiores a 7 se pueden obtener de (3) suprimiendo pares de categorías opuestas.

En general, las estructuras de escala (1) y (3) pueden extenderse hasta la escala verbal bipolar de 15 puntos añadiendo categorías con modificadores el (*extra bajo*), *vl* (*muy bajo), vh (muy alto), eh (extra alto)* Para el adjetivo *p* y su *op* de la siguiente manera:

L= (ehap, vhap, hap, map, lap, vlap, elap, np, elp, vlp, lp, mp, hp, vhp, ehp). ……..(4)

Las escalas bipolares con el número de categorías inferiores a 15 se pueden obtener de (4) suprimiendo pares de categorías opuestas. A continuación se muestra un ejemplo de tal escala bipolar de 9 puntos:

L = (extremadamente a disgusto, muy a disgusto, no le gusta moderadamente, no le gusta poco, ni gusta ni disgusta, gusta ligeramente, gusta moderadamente, gusta mucho, gusta extremadamente),

Con el centro C = ni cómo ni desagrado, y con la estructura reducida de (4):

L= (*ehap, vhap, map, lap, np, lp, mp, vhp, ehp*).

Un ejemplo de escala bipolar verbal de 13 puntos se puede encontrar en (Gunderman, 2013).

Las estructuras consideradas son comunes para las escalas utilizadas en sociología, psicología y medicina, pero en la minería de opinión y en los sistemas de recomendación, las escalas bipolares con simetría en categorías opuestas pueden tener categorías expresadas en una variedad de formas. Debajo se encuentra la escala de Ringo para calificar música (Shardanand, 1995) en su sistema de recomendación musical que explícitamente es

diferente de las estructuras (1) y (3) pero implícitamente tiene la forma simétrica como en (1) con el centro en la categoría 4:

L = (1. Pass the earplugs. 2. Barely tolerable. 3. Eh. Not really my thing. 4. Doesn't turn me on, doesn't bother me. 5. Good Stuff. 6. Solid. They are up there. 7. BOOM! One of my FAVORITE few! Can't live without it)

Las estructuras típicas consideradas de las escalas bipolares pueden utilizarse para la construcción de escalas verbales bipolares con simetría en categorías opuestas. Las etiquetas verbales de estas categorías pueden diferir de una aplicación a otra, pero la estructura simétrica debe ser preservada para considerar las escalas verbales como escalas bipolares.

En la siguiente sección, daremos la definición formal de una escala bipolar finita como un conjunto ordenado de índices de categorías de la escala bipolar considerada anteriormente. A continuación, consideraremos las funciones de puntuación bipolar que asignan los valores numéricos a cada punto de la escala bipolar que se utilizará como modelo numérico de esta escala. Las principales propiedades de las funciones de utilidad bipolar estarán relacionadas también con la operación de negación relacionada con la simetría en las puntuaciones asignadas a categorías opuestas de la escala bipolar.

3.6 Escalas bipolares finitas

Formalmente la escala bipolar L con n categorías ordenadas $c_1 < \ldots < c_n$ puede ser representada por el conjunto ordenado de índices de estas categorías $J = \{1, \ldots, n\}$, $n > 1$, con la operación de *negación* N: $J \rightarrow J$ definida por

$N(j) = n + 1 - j$ for all $j \in J$. …… (5)

La negación (5) es función *estrictamente decreciente*:

$N(i) > N(j)$ if $i < j$,……...(6)

Y *es involutivo*:

$N(N(j)) = j$, for all $j \in J$..……...……(7)

Además, supongamos que la escala bipolar tiene un número impar de elementos, es decir, $n = 2m + 1$ para un cierto número entero positivo m. En este caso, el conjunto J tendrá el *punto fijo de la negación N*, i.e., un elemento C tal que:

$N(C) = C$. ……...……...(8)

La propiedad (8) se cumple para el elemento único

$C = m+1$, ……...(9)

Que se llama centro, neutro o punto medio de la escala bipolar. De (5) y (6) se deduce que la escala bipolar con número par de elementos n no tiene centro. Las escalas bipolares sin centro se pueden obtener de las escalas con el centro borrando el centro.

Por ejemplo, la escala bipolar de 5 puntos *(nunca, rara vez, a veces, a menudo, siempre)* puede ser dada por el conjunto ordenado de índices $J = \{1, 2, 3, 4, 5\}$ y con la negación $N(j) = 6 - j$, tal que $N(1) = 5$, $N(2) = 4$, $N(3) = 3$, $N(4) = 2$, $N(5) = 1$. Esta escala tiene el centro $C = 3$.

Para (5), $n = 2m+1$ y (9) obtenemos: $N(j) + j = 1 + n$, y

$N(j) + j = 2C$, for all j in J. (bipolarity).....(10)

Debido a la bipolaridad, los elementos $N(j)$ y j de la escala están simétricamente situados con respecto al centro y los "polos" 1 y n de la escala J:

$N(j) - C = C - j$, $N(j) - 1 = n - j$, $j - 1 = n - N(j)$, for all $j \in J$.

El conjunto ordenado $K = \{-m, ..., -1, 0, 1, ..., m\}$ será considerado como la forma centrada de la escala bipolar $J = \{1, \ldots, 2m + 1\}$, $m > 0$. La operación de negación N: K → K en K se definirá por:

$N(k) = -k$, for all $k \in K$….....(11)

Si no es engañosa, usaremos la misma letra N para las negaciones en J y en K usando argumentos j o k respectivamente. Es claro que N sobre K es una función decreciente e involutiva, i.e., $N(N(k)) = k$, para todo $k \in K$. Esta escala tiene el centro $C = 0$ con $N(C) = C = 0$. De (11) sigue la bipolaridad (10) para la escala K con la negación (11):

$N(k) + k = 2C$, for all $k \in K$……...(12)

Además, la escala K = {-m, ..., m}, m> 0, con la negación definida por (11) se denominará escala bipolar centrada.

Por ejemplo, la escala bipolar de 5 puntos J = {1, 2, 3, 4, 5} se representará en forma centrada como: K = {-2, -1, 0, 1, 2} con la negación N sobre K definido por $N(-2) = 2$, $N(-1) = 1$, $N(0) = 0$, $N(1) = -1$, $N(2) = -2$ y con el centro $C = 0$.

Las escalas bipolares J = {1, ..., 2m + 1}, m> 0 y K = {-m, ..., m} pueden transformarse una a otra de la siguiente manera:

$k = j - m - 1$, $j = k + m + 1$, for all j in J and for all k in K. …….(13)

3.7 Funciones de puntuación bipolares

Aquí estudiamos las propiedades de las funciones de puntuación (utilidad) definidas en las escalas bipolares. Las funciones de puntuación dan la posibilidad de modelar usuarios con diferentes utilidades de categorías de la misma escala. Por ejemplo, un usuario prefiere usar las categorías de la escala cerca de los polos, pero otro usuario prefiere utilizar las categorías cerca del centro. Para tales usuarios la utilidad de las mismas categorías puede ser diferente. También en el desarrollo de sistemas de toma de decisiones o de recomendación usando escalas bipolares la utilidad de categorías de las escalas puede ser representada por funciones no lineales que modelan la utilidad de categorías de diferente manera dependiendo de la aplicación o de la tarea.

La utilidad de las clasificaciones de la escala bipolar puede ser representada por funciones no lineales. Para tales funciones, la diferencia entre las clasificaciones vecinas depende de sus posiciones en la escala. Por ejemplo, para la escala bipolar *(nunca, rara vez, a veces, a menudo, siempre)* la diferencia entre la utilidad de las categorías a veces ya menudo puede ser modelada por el número 10, pero la diferencia entre las utilidades de las

categorías a menudo y siempre puede ser Modelado por el número 50. Tales funciones de utilidad no lineal pueden utilizarse, por ejemplo, para modelar las calificaciones de los usuarios en el enfoque basado en modelos para el filtrado colaborativo en sistemas de recomendación (Adomavicius, 2005).

En el enfoque basado en modelos para la construcción de funciones de utilidad bipolar, no nos preocuparemos mucho acerca de cómo el usuario mide estas utilidades; Vamos a pensar cómo modelar estas utilidades para diferentes usuarios y diferentes tareas. En el enfoque basado en modelos, en lugar de la medición "adecuada" de las preferencias o calificaciones del usuario será más importante la efectividad de las recomendaciones o decisiones obtenidas en la salida del recomendador o del sistema de toma de decisiones utilizando estas funciones de utilidad bipolar.

Generalmente, estas funciones de utilidad pueden ser parametrizadas y ajustadas o ajustadas por algún procedimiento de aprendizaje automático con el objetivo de obtener las recomendaciones y decisiones óptimas o útiles sobre la salida del sistema. Lo que requeriremos de estas funciones de utilidad es que necesitan satisfacer las propiedades similares a las propiedades de las escalas bipolares que consideramos antes. En esta sección, consideraremos estas propiedades de las funciones de utilidad bipolar y en la siguiente sección consideraremos diferentes métodos de construcción de tales funciones.

Además, denotaré cualquiera de las escalas bipolares J o K con $n = 2m+1$, $m > 0$, categorías para $J = \{1, \ldots, 2m+1\}$ tenemos $C = m+1$, $N(j) = n+1-j$ para todos los j in J *y para la escala* $K = \{-m, \ldots, -1, 0, 1, \ldots, m\}$ tenemos $C = 0$, $N(k) = -k$, *para todos los* k in K. para ambas escalas tenemos $N(C) = C$. Llamamos a los valores P1 = min (I) $P_1 = \min(I)$ y $P_2 = \max(I)$ los polos negativo y positivo, correspondientemente. Para la escala J, tenemos $P_1 = 1$, $P_2 = n$, y para la escala K tenemos $P_1 = -m$, $P_2 = m$. Para ambas escalas, tenemos $N(P_1) = P_2$ and $N(P_2) = P_1$.

Definición 1. Sea I una escala bipolar con negación N. Una función real estrictamente creciente $U: I \rightarrow R$ se llamará función de puntuación o de utilidad en I. Esta función se llamará *función de puntuación bipolar* (BSF) en I si se cumple la siguiente condición:

$U(N(i)) + U(i) = U(P_1) + U(P_2)$, para todos los $i \in I$. (bipolarity) ……(14)

Un BSF se llamará una función de puntuación bipolar centrada (CBSF) si $U(C)= 0$.

Para las escalas con el centro C de N (C) = C, tenemos U (P1) + U (P2) = 2U (C) y la propiedad de la bipolaridad puede ser dada por:

$U(N(i)) + U(i) = 2U(C)$, for all $i \in I$. (bipolarity)……(15)

De (14) y (15) tenemos:

$|U(j) - U(C)| = |U(N(j)) - U(C)|$, $|U(j) - U(P_1)| = |U(N(j)) - U(P_2)|$, for all $j \in J$,

Es decir, para la función de puntuación bipolar U, los valores de utilidad de las categorías opuestas $U(j)$ y $U(N(j))$ están a igual distancia del valor de utilidad de la categoría neutra $U(C)$ y la misma distancia de los valores de utilidad de los polos.

A partir de la definición de la función de puntuación bipolar centrada, se sigue:

$U(N(i)) = -U(i)$, for all $i \in I$..……...(16)

Para el BSF centrado en la escala bipolar centrada $K= (-m, \ldots, m)$, $m > 0$, tenemos:

$U(0) = 0$, $U(k) > 0$ if $k > 0$, $U(k) < 0$ if $k < 0$, y

$U(-k) = -U(k)$, para todos los $k \in K$…..……...…(17)

Las funciones de puntuación bipolares centradas dan los modelos naturales de utilidad de categorías de escalas bipolares de calificación verbal cuando estas categorías tienen sentimientos positivos y negativos. Por ejemplo, para la escala bipolar centrada de 5 puntos K = {-2, -1, 0, 1, 2}, se puede definir la función de puntuación bipolar centrada $U(K)$ = {-10, -4, 0, 4 , 10} preservando el signo y la simetría de la escala bipolar K.

Proposición 1. Si U es una función de puntuación bipolar en I, entonces la función $W: I \to \mathbb{R}$ definida por

$W(i) = pU(i) + q$ for all $i \in I$, ... (18)

Donde $p, q \in \mathbb{R}$, $p > 0$, Son números reales, será también la función de puntuación bipolar en I.

Proof. Está claro que W es estrictamente creciente. La bipolaridad de W se deduce de (18) y la bipolaridad (15) de U: $W(N(i))+ W(i) = pU(N(i)) + q + pU(i) + q = p(U(N(i)) + U(i))+2q = p(2U(C))+2q = 2(pU(C)+q) = 2W(C)$

En (18) los parámetros p y q definen el escalamiento y desplazamiento de la función de utilidad U correspondientemente.

De la Proposición 1 se deduce que de cualquier función de puntuación bipolar U se puede obtener la función de calificación bipolar centrada U_C como sigue:

$U_C(i) = U(i) - U(C)$ for all $i \in I$. ……(19)

De (5) y (9) se deduce que la función de identidad: $U(j) = j$, para todo $j \in J$, será la función de puntuación bipolar. Esta función se denominará función de puntuación bipolar estándar (SBSF). Por ejemplo, SBSF definido en la escala bipolar de 7 puntos J = {1, 2, 3, 4, 5, 6, 7} tendrá los valores $U(J) = J$ = {1, 2, 3, 4, 5, 6, 7}. A partir de esta función, aplicando la transformación (18) con valores de parámetro p = 1, q = -1, podemos obtener otra función de puntuación bipolar popular $U(J)$ = (0, 1, 2, 3, 4, 5, 6).

Como se desprende de los últimos ejemplos, la mayoría de las escalas de calificación populares, incluidas las escalas de Likert (1932) y las escalas utilizadas en los sistemas de recomendación Shardanand(1995) utilizan funciones de puntuación bipolares estándar. En este trabajo, nos interesó principalmente con funciones de puntuación bipolar no lineal. Por ejemplo, en la escala bipolar de 7 puntos J = {1, 2, 3, 4, 5, 6, 7} con el centro $C = 4$, la función de puntuación bipolar no lineal puede definirse como sigue: $U(J)$ = {0 (15), 15, 40, 50, 60, 85, 100} con $U(C)$ = 50. A partir de esta BSF aplicando la transformación lineal (19) se puede obtener la función de puntuación bipolar centrada $UC(J)$ = {-50, -35 , -10, 0, 10, 35, 50}.

Obsérvese que mediante transformaciones (19) de los índices de las escalas bipolares J = {1, ..., 2m + 1} y K = {-m, ..., m} se puede transformar BSF U_J: $J \to R$ definida en J en BSF U_K: $K \to R$ definido en K y viceversa como sigue:

$U_J(j) = U_K(j - m - 1)$, $U_K(k) = U_J(k + m + 1)$, for all $j \in J$ and $k \in K$. (20)

Ambas funciones tendrán el mismo conjunto de valores: *UJ(J)* = *UK(K)*. Por ejemplo, el CBSF *UK(K)* = {-10, -4, 0, 4, 10} definido en *K* = {-2, -1, 0, 1, 2} se transformará en CBSF *UJ(J)* = { -10, -4, 0, 4, 10} definido en *J* = {1, 2, 3, 4, 5}. Esta transformación simple se puede utilizar en el método de construcción de CBSF en *J* considerado en la siguiente sección: definir CBSF por alguna función impar en *K* y luego cambiar en esta función el conjunto de argumentos (índices) *K* por *J*.

4. METODOLOGÍA

Para la clasificación de la polaridad utilizando sn-gramas se realizaron los siguientes pasos metodológicos:

- Selección del corpus.
- Análisis sintáctico.
- Algoritmo para la extracción de bigramas y trigramas sintácticos.
- Creación de la matriz termino documento para bigramas y trigramas sintácticos.
- Selección de clasificadores.
- Generación del conjunto de datos para entrenamiento y evaluación.

Los pasos metodológicos se describen a continuación.

4.1 Selección del corpus

El corpus utilizado es muchocine, este corpus contiene 3878 reseñas de películas recolectadas del sitio web MuchoCine (Cruz 2008), aproximadamente 2 millones de palabras, con una media de 546 palabras por crítica (Cruz, 2008).

Las reseñas son escritas por los usuarios del sitio web. Las reseñas son calificadas en un rango del 1 al 5, algunos autores como Martín (2013) usa las reseñas marcadas como 1 y 2 como negativas y las reseñas marcadas con 4 y 5 positivas.

Tabla 1. Distribución de reseñas por rango

Rango	**Número de Reseñas**
1	351
2	923
3	1253
4	890
5	461
Total	**3878**

En la tabla 1 se puede observar la distribución de del rango de calificación de las frecuencias siendo el número tres el de mayor número. Para uno de los conjuntos de datos de entrenamiento y evaluación el número tres se considera como neutral con el valor 0, al igual que Martin(2013) el uno y dos se toman como negativos con el valor de -1 y el cuatro y cinco como positivos con el valor de 1.

Las reseñas no cuentan con la misma extensión y al ser escritas por los usuarios del sitio web contienen errores ortográficos, la mayoría de las reseñas son escritas por diferentes autores aunque existen casos en los que un autor critica más de una película. El corpus se encuentra en idioma español.

La idea de utilizar este corpus se debe a que además de las reseñas el corpus contiene los arboles de dependencia de cada reseña y los arboles de dependencia de los resúmenes de cada reseña, los arboles de dependencia fueron realizados con el analizador sintáctico FreeLing en la versión 1.3. El contar con el análisis sintáctico facilita la extracción de los sn-gramas sintácticos que se utilizarán como rasgos.

También es importante comentar que la realización de trabajos enfocados a la clasificación de la polaridad en idioma español es escasa siendo que el idioma español es uno de los idiomas con más presencia en internet, sin embargo se pretende que la propuesta planteada se generalice y pueda aplicarse a otros idiomas.

4.2 *Análisis sintáctico*

Los sn-gramas que se utilizarán como características para la etapa de clasificación se obtienen a partir de un árbol de dependencias, el corpus contiene en archivos de texto separados los arboles de dependencia generados con FreeLing en la versión 1.3 (Atserias, 2006) tanto de las críticas como de sus resúmenes dando un total de 7,756 archivos. Con esto se ahorra el trabajo de la generación del análisis sintáctico y se puede trabajar en otras tareas.

```
sn/top/(Bodrio bodrio NCMS000 -) [
  sp-de/mod/(de de SPS00 -) [
    sn/gen/(terror terror NCMS000 -) [
      grup-sp/modnomatch/(a a SPS00 -) [
        sn/head/(francesa francés AQ0FS0 -) [
          j-fs/modnomatch/(la el DA0FS0 -)
        ]
      ]
    ]
  ]
]
```

Figura 1. Árbol de dependencia del resumen de la crítica número 7 del corpus muchocine

En la figura 1 se muestra un ejemplo de un árbol de dependencia del corpus muchocine el cual contiene etiquetas sintácticas y morfológicas, el orden en el que aparecen es {synt/func/(form lemma tag-)}, cada una de estas líneas es un nodo().

El símbolo "/" sirve para separar las etiquetas, los símbolos "(,-,)" encierran la etiqueta morfológica **"tag",** los símbolos "[,]" representan la jerarquía del árbol.

4.3 *Algoritmo para la extracción de bigramas y trigramas sintácticos*

Para utilizar los sn-gramas como rasgos de los vectores de entrenamiento previamente deben extraerse del árbol de dependencia para lo cual se creó un algoritmo el cual fue implementado posteriormente en el lenguaje de programación de python 2.7 que permite extraer de manera automática bigramas continuos de palabras, bigramas continuos de lemas, trigramas continuos de palabras y trigramas continuos de lemas.

Los sn-gramas se extraen de cada árbol, de cada archivo que contiene el análisis sintáctico de las críticas. El análisis sintáctico contiene más de un árbol de dependencias uno por cada frase u oración, el número de árboles de cada archivo varía de acuerdo a la extensión de la crítica o el resumen.

Lo primero que el algoritmo detecta es la raíz del árbol la cual es representada por la func top en la estructura de FreeLing, el algoritmo para la obtención de la raíz del árbol es el siguiente.

```
1 Obtener la sangría de la línea.
2 Si la línea del archivo contiene “/top/”:
3         raiz=obtenerPalabra(línea,sangría).
4 Agregar la raíz a una lista para saber el número de árboles de la crítica.
```

La raíz indica que empieza un nuevo árbol de dependencia dentro de los archivos del análisis sintáctico de la crítica. También es importante la sangría de cada línea del árbol ya que junto con los caracteres [y] determinan la jerarquía la cual incrementa el sangrado en 2 dependiendo del nivel jerárquico. Una vez que se tiene la raíz se sabe que su sangría es 0 y se pasa a la siguiente línea del árbol donde se obtiene la primer palabra y se concatena con la raíz para obtener los primeros bigramas sintácticos que corresponden de la unión de la raíz con los nodos siguientes del árbol, para el árbol de ejemplo de la figura 1 la raíz sería “Bodrio” y la segunda línea arrojaría la palabra “de” es decir el primer bigrama sintáctico sería “Bodrio_de” según el siguiente pseudocódigo del algoritmo.

```
Si sangría==2 y el carácter "]" no se encuentra en la línea:
        palabra=obtenerPalabra(línea,sangria).
        bigrama=raiz+ "_" +palabra.
        obtenemos la frecuencia del bigrama en la crítica.
        palabrasAnt[sangria]=palabra
```

El mismo proceso se realiza de manera recursiva con los demás niveles jerárquicos del árbol de dependencias tomando en cuenta la sangría y los caracteres [y], los bigramas sintácticos resultantes serian Bodrio_de, de_terror, terror_a, a_francesa y francesa_la derivados del árbol de dependencia de la figura 1.

```python
def obtenerBigramas(self,archivo):
    self.dicLemasArchivo.clear()
    palabra=""
    palabra2=""
    sangria=0
    root=""
    contador=4

    for lineas in archivo:
        print "Extracting data, please wait..."
        sangria=self.extraer.obtenerNivel(lineas)
        if "/top/" in lineas:
            root=self.extraer.obtenerPalabra(lineas,sangria)
            self.extraer.raiz.append(root)

        if sangria==2 and "]" not in lineas:
            palabra=self.extraer.obtenerPalabra(lineas,sangria)
            bigrama= root+"_"+palabra
            frecuencia=self.obtenFrecuencia(bigrama)
            self.dicLemasArchivo[bigrama]=frecuencia
            self.extraer.palabrasAnt[sangria]=palabra

        if sangria==contador and "]" not in lineas and sangria!=0 and sangria!=2:
            palabra3=self.extraer.obtenerPalabra(lineas,sangria)

            if palabra2==="":
                palabra2=self.extraer.palabrasAnt[sangria-2]
            bigrama2=palabra2+"_"+palabra3
            frecuencia=self.obtenFrecuencia(bigrama2)
            self.dicLemasArchivo[bigrama2]=frecuencia
            self.extraer.palabrasAnt[sangria]=palabra3

        if "[" in lineas and sangria!=0:
            palabra2=self.extraer.obtenerPalabra(lineas,sangria)
            contador=sangria+2

        if "]" in lineas:
            palabra2=""
            contador=sangria
```

Figura 2. Implementación del algoritmo de extracción en python.

La figura anterior muestra el código de implementación en python del algoritmo a través del método obtenerBigramas el cual se encuentra dentro de la clase ExtraerBigramas. Este es el código base para extraer bigramas sintácticos de palabras pero se utiliza también para la obtención de bigramas sintácticos de lemas.

```
def obtenerPalabra(self,linea,sangria):
    palabra=""
    posicionFinal=0
    self.caracterPorLinea(linea)
    posicionInicial=self.caracterLinea.index("(")+1
    if sangria==0:
        posicionFinal=self.caracterLinea.index(" ")
    if sangria > 0:
        for x in range(posicionInicial,len(self.caracterLinea)):
            if self.caracterLinea[x]==" ":
                posicionFinal=x
                break
    if "Fc(" in linea or "VMN0000(" in linea or "SPS00(" in linea:
        posicionFinal=self.caracterLinea.index(")")

    palabraLista=self.caracterLinea[posicionInicial:posicionFinal]
    for x in palabraLista:
        palabra=palabra+x
    del self.caracterLinea[0:len(self.caracterLinea)]
    return palabra
```

Figura 3. Implementación del método obtenerPalabra de la clase Extracción.

La validación de la extracción se comparó extrayendo manualmente los bigramas de 28 árboles de dependencia de diferente extensión y críticas con lo cual se corrigieron los errores observados garantizando que la extracción es correcta en arboles de dependencia de árboles jerárquicamente cortos y extensos con un nivel de profundidad alto.

Falta realizar la documentación adecuada de los programas para colocarlos como recursos abiertos para que faciliten la extracción de sn-gramas de los árboles de dependencia en formato de FreeLing para otras tareas donde puedan aplicarse.

```python
def obtenerLema(self,linea,sangria):
    lema=""
    posicionFinal=0
    posicionInicial=0
    posicionPivote=0
    self.caracterPorLinea(linea)

    if sangria==0:
        indiceUno=0
        indiceDos=0
        for x in self.caracterLinea:
            if x==" ":
                posicionInicial=indiceUno+1
                break
            indiceUno=indiceUno+1

        for x in range(posicionInicial,len(self.caracterLinea)):
            if self.caracterLinea[x]==" ":
                #posicionFinal=indiceUno+indiceDos
                posicionFinal=(indiceDos+1)+posicionInicial
                #indice=0
                break
            indiceDos=indiceDos+1

    if sangria > 0:
        indiceUno=0
        indiceDos=0
        indiceTres=0
        for x in self.caracterLinea:
            if x!=" ":
                posicionPivote=indiceUno

                break
            indiceUno=indiceUno+1
        for x in range(posicionPivote,len(self.caracterLinea)):
            if self.caracterLinea[x]==" ":
                posicionInicial=indiceDos+posicionPivote+1

                break
            indiceDos=indiceDos+1
        for x in range(posicionInicial,len(self.caracterLinea)):
            if self.caracterLinea[x]==" ":
                posicionFinal=(indiceTres+1)+posicionInicial

                break
            indiceTres=indiceTres+1

    lemaLista=self.caracterLinea[posicionInicial:posicionFinal]


    for x in lemaLista:
        lema=lema+x

    del self.caracterLinea[0:len(self.caracterLinea)]
    lema=lema.strip()
    return lema
```

Figura 4. Implementación del método obtenerLema de la clase Extracción.

4.4 *Creación de matriz termino documento*

Con los sn-gramas se realizó la matriz termino documento donde por una parte se tienen las 3878 criticas que son los documentos y por otra los bigramas y trigramas sintácticos de palabras y lemas. Como rasgos o características y el valor de cada uno es su medida *tf* y un conjunto con pesos en *tfidf*.

Se generaron ocho matrices una para bigramas sintácticos de palabras una con pesos *tf* y otra pesos *tfidf*, bigramas sintácticos de lemas una con pesos *tf* y otra con pesos *tfidf*, trigramas sintácticos de palabras una con pesos *tf* y otra con pesos *tfidf*, trigramas sintácticos de lemas una con peso *tf* y otra con pesos *tfidf*.

En este semestre se generaron además las matrices término documento con la generación del análisis sintáctico de las críticas en la versión 3.1 de FreeLing para el mismo conjunto de datos bigramas de palabras, bigramas de lemas, trigramas de palabras y trigramas de lemas utilizando como pesos *tfidf*.

La generación de las matrices se realiza de manera automática con la programación en python de un algoritmo que guarda la matriz tanto en Excel como en un archivo de texto.

```python
import xlsxwriter

class CreacionCVS:
    diccionarioFrecuencias={}
    libro=xlsxwriter.Workbook('TrigramasLemasBueno.xlsx')
    ws1=libro.add_worksheet()
    ws1.write(0,0, 'Id')
    ws1.write(0,1, 'Bigrama')
    contadorBigrama=1

    def insetarBigrama(self,bigrama,documento,frecuencia,numSngramas,nombreDocumento):
        tf=frecuencia/float(numSngramas)
        nombre=self.obtenerNombre(nombreDocumento)

        if bigrama in self.diccionarioFrecuencias:
            y=self.diccionarioFrecuencias[bigrama]
            self.ws1.write(0,documento+1,nombre)
            self.ws1.write(y,0,y)
            self.ws1.write(y,1,bigrama)
            self.ws1.write(y,documento+1,tf)
        else:
            self.ws1.write(0,documento+1,nombre)
            self.ws1.write(self.contadorBigrama,0,self.contadorBigrama)
            self.ws1.write(self.contadorBigrama,1,bigrama)
            self.ws1.write(self.contadorBigrama,documento+1,tf)
            self.diccionarioFrecuencias[bigrama]=self.contadorBigrama
            self.contadorBigrama=self.contadorBigrama+1

    def guardarLibro(self):
        self.libro.close()
```

Figura 7. Clase CreacionCVS que genera la matriz termino documento en formato xlsx.

La matriz para los bigramas sintácticos de palabras es del orden de 642456*3878, la matriz de bigramas sintácticos de lemas es del orden de 486261 * 3878 y la matriz de trigramas sintácticos de lemas tiene un tamaño de 919733 * 3878.

```
class CreacionTxt:
    diccionarioFrecuencias={}
    a = open('TrigramasLemasBueno.txt', 'w')
    contadorBigrama=1

    def insetarBigrama(self,bigrama,documento,frecuencia,numSngramas,nombreDocumento):
        tf=frecuencia/float(numSngramas)
        nombre=self.obtenerNombre(nombreDocumento)
        self.a.write('\n'+ bigrama)
        self.a.write('/'+ nombre)
        self.a.write('/%f' %tf)


    def obtenerNombre(self,nombreDocumento):
        caracterLinea=[]
        posicionInicial=0
        indiceUno=0
        indiceDos=0
        nombre=""
        for i in nombreDocumento.lower():
            caracterLinea.append(i)

        for x in caracterLinea:

            if x=="\\":
                posicionInicial=indiceUno+1
                break
            indiceUno=indiceUno+1

        for x in range(posicionInicial,len(caracterLinea)):
                if caracterLinea[x]==".":
                    #posicionFinal=indiceUno+indiceDos
                    posicionFinal=indiceDos+posicionInicial
                    #indice=0
                    break
                indiceDos=indiceDos+1
        nombreLista=caracterLinea[posicionInicial:posicionFinal]

        for x in nombreLista:
            nombre=nombre+x
        del caracterLinea[0:len(caracterLinea)]
        return nombre
```

Figura 6. Clase CreacionTxt que genera la matriz termino documento en formato txt.

4.5 *Selección de clasificadores*

Para la tarea de clasificación se seleccionaron los clasificadores NaiveBayes y Support Vector Machine debido a que estos clasificadores son los que obtienen un mayor grado de precisión en la tarea de clasificación de polaridad y son utilizados por Martín (2013) en un estudio previo sobre el mismo corpus reportando de igual manera los resultados más altos . Para su uso se utilizara la herramienta WEKA.

Se utilizó la función SMO de weka con la función supportVector PolyKernel, también se utilizó la librería libSVM Chang(2011).

4.6 *Generación del conjunto de datos para entrenamiento y evaluación*

Para utilizar la herramienta WEKA se generaran los archivos en formato arff para los cuales se extraen de manera automática de la matriz de término-documento creada previamente tomando cada documento como un vector de datos con la medida ***tf*** y otro con la medida ***tfidf*** de cada atributo para este trabajo los sn-gramas.

Se generaron archivos sparce arff para quitar la dispersión de los datos es decir no tomando en cuenta aquellos atributos que tienen 0.0 en la frecuencia disminuyendo con esto el tamaño del archivo arff, la extracción y generación del archivo se realizó de manera automático con la programación de una clase para este fin en python utilizando la librería arff y su método dump.

Se generaron dos programas uno para extraerlos de la matriz termino documento en Excel y otro para la matriz en texto. También se generó una clase más para marcar cada vector con su clase correspondiente en este caso el ranking de cada crítica que va del 1 al 5. También se generó un conjunto de datos con polaridades negativas marcadas con -1 conteniendo las críticas con calificación uno y dos, neutras marcadas con 0 con la calificación de tres y positivas marcadas con 1 con calificaciones cuatro y cinco.

4.7 *Análisis sintáctico de las críticas del corpus*

Para la extracción de los sn-gramas como menciona Sidorov (2013).se requiere de un análisis sintáctico previo una razón como se mencionó previamente para la selección del corpus utilizado es que este contiene dicho análisis sintáctico de cada una de las criticas sin embargo este análisis se realizó con la versión 1.3 del analizador sintáctico de Freeling (Atserias, 2006). El cual actualmente se encuentra en la versión 3.1.

Se optó por generar el análisis sintáctico de las críticas con la versión actual de Freeling (Atserias, 2006).

Para este proceso se utilizó el recurso denominado “FreeLing Dependency parser Web Service v.3” del proyecto PANACEA Bel (2012). Se obtuvieron los archivos con los arboles de dependencias de cada una de las críticas del corpus con la versión 3.1 del analizador sintáctico de Freeling (Atserias, 2006).

```
import arff
import glob,os
from ClaseExtraccion import Extraccion
from ClasePolaridad import Polaridad
data=[]
temporal=[]
nombres=[]
#nombresDos=[]
documentos=[]
diccionario={}
a = open('TrigramasBueno.txt', 'r')
extraer=Extraccion(a)
pol=Polaridad()
pol.obtenerRank()

def extraerValores(a):
    contador=0
    nombres.append("critica")

    for linea in a:
        bigrama=extraer.bigramaArff(linea)
        documento=extraer.documentoArff(linea)
        frecuencia=extraer.frecuenciArff(linea)
        llave=bigrama.strip()+"*"+documento.strip()
        diccionario[llave]=frecuencia
        if(bigrama not in nombres):

            nombres.append(bigrama)

        if(documento not in documentos):
            documentos.append(documento)

def generarDatos():
    documento=0
    for x in documentos:
        del temporal[0:len(temporal)]
        temporal.append(x)
        for y in nombres:
            frecuencia=obtenerFrecuencia(x,y)
            frecuencia=float(frecuencia)
            if 'critica' != y:
                temporal.append(frecuencia)


        temporal.append(pol.polaridad(x))
        print (len(temporal))
        data.append(list(temporal))
        break;
```

```
def obtenerFrecuencia(x,y):
    frecuencia=0
    documento=x.strip()
    bigrama=y.strip()
    llave=bigrama+"*"+documento
    if diccionario.has_key(llave):
        frecuencia=diccionario[llave]
    else:
        frecuencia=0.0

    return frecuencia

extraerValores(a)
generarDatos()

arff.dump('trigramasAtributos.arff',data,relation="Polaridad",names=nombres)
```

Figura 8. Programa para generar

4.8 *Métodos de construcción de las funciones de puntuación bipolar*

Como se desprende de la sección anterior, la función de puntuación bipolar en una escala bipolar es una transformación lineal de una función impar con un valor de argumento cero en el centro de la escala. La mayoría de las escalas de calificación tradicionales pueden considerarse como casos particulares de funciones de puntuación bipolar cuando la función de puntuación es lineal. En esta sección, consideramos diferentes métodos heurísticos de construcción de funciones de utilidad bipolar (puntuación) que son, generalmente, no lineales.

4.8.1 Funciones de utilidad bipolares basadas en la distribución de las calificaciones de usuario

Considere un método heurístico de construcción de funciones de utilidad bipolar basado en la distribución de las puntuaciones bipolares obtenidas de las calificaciones del usuario. Para simplificar, considere las categorías en una escala bipolar centrada K = (-m, ..., -1, 0, 1, ..., m). Supongamos que el usuario ha realizado valoraciones de N elementos con frecuencias $P = (P_{-m}, \ldots, P_{-1}, P_0, P_1, \ldots, P_m)$ con $P_{-m} + \ldots + P_m = N$, donde P_k, $k \in \{-m, \ldots, m\}$ es la frecuencia de la categoría de la escala K. El algoritmo propuesto tiene los siguientes pasos.

1. Frecuencias de simetría:

$$PS_k = \frac{P_k + P_{-k}}{2}, \text{ for all } k = -m, \ldots, m.$$

2. Calcule la función de distribución acumulativa:

$U_{-m} = 0, \quad U_{k+1} = U_k + c|PS_{k+1} - PS_k| + d, \qquad k = -m, \ldots, m-1,$.....(21)

Donde c es alguna constante de escala y $d = 0$ if $PS_{k+1} \neq PS_k$ for all $k = -m, \ldots, m-1$, de otra forma $d > 0$ es igual a alguna pequeña constante no negativa. La constante d se añade para asegurar que todas las categorías tendrán puntuaciones diferentes y la función de utilidad será estrictamente creciente.

3. Normalizar la función de utilidad para obtener los valores de función de utilidad en el rango $[0, M]$, or $[-M, M]$, $M > 0$ y, si es necesario, moverlo desde la escala $K = \{-m, \ldots, m\}$ a la escala $J = \{1, \ldots, 2m + 1\}$ sustituyendo índices por (20).

Se puede demostrar que el método propuesto construye una función de puntuación bipolar que satisface la propiedad de bipolaridad.

La idea detrás del método propuesto para la construcción de la función de utilidad bipolar es la siguiente: "Cuanto mayor sea la diferencia entre las frecuencias de dos categorías vecinas, mayor será la diferencia en las utilidades de estas categorías".

La figura 9 muestra un ejemplo sintético de construcción de funciones de utilidad bipolares basadas en la distribución P = (5, 35, 48, 60, 30, 20, 2) de categorías de escala bipolar de 7 puntos en N = 200 calificaciones de usuario. A la izquierda están las distribuciones originales y simétricas, y ahí está la función de utilidad bipolar construida a partir de la distribución simetrizada. En (21) se utiliza d = 0 porque en la distribución simetrizada dos categorías vecinas tienen frecuencias diferentes.

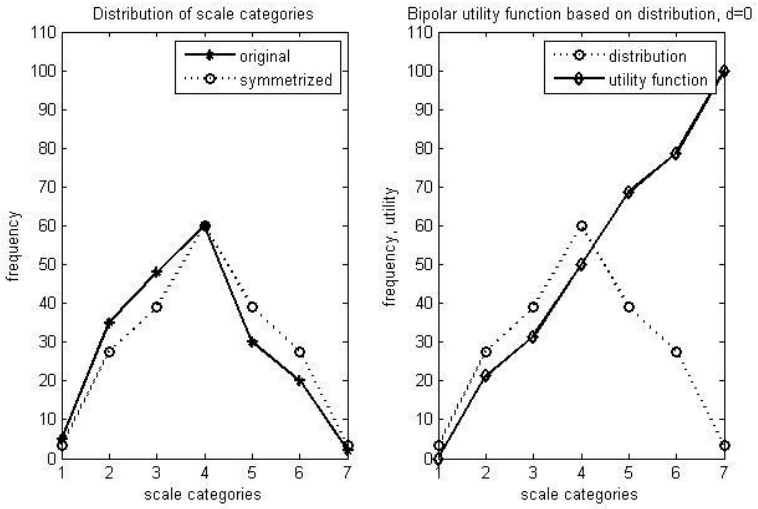

Figura 9. Construcción de la función de utilidad bipolar basada en la distribución simétrica de las categorías de las escalas bipolares obtenidas en N = 200 puntuaciones de usuarios, d = 0

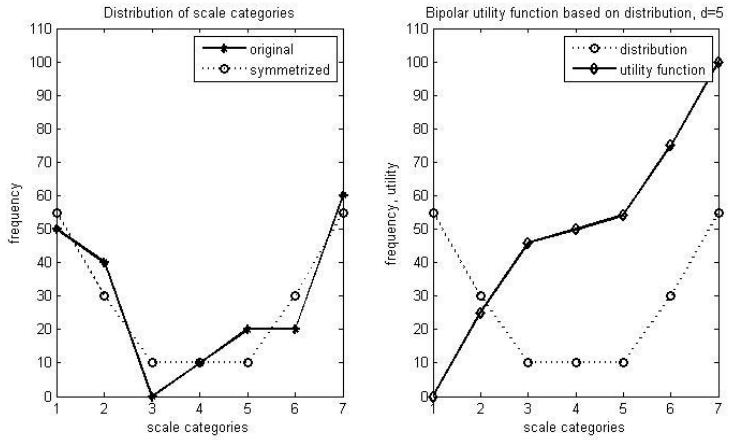

Figura 10. Construcción de la función de utilidad bipolar basada en la distribución simétrica de las categorías de las escalas bipolares obtenidas en N = 200 puntuaciones del usuario, d = 5.

La figura 2 representa un ejemplo sintético de construcción de funciones de utilidad bipolares basadas en la distribución P = (50, 40, 0, 10, 20, 20, 60) de categorías de escala bipolar de 7 puntos en N = 200 puntuaciones de usuario. A la izquierda se encuentran las distribuciones originales y simétricas. Justo allí es la función de utilidad bipolar construida a partir de la distribución simetrizada. En (21) se utiliza d = 5 porque en la distribución

simetrizada dos categorías vecinas tienen frecuencias iguales. En ambos casos usamos c = 1. Todas las funciones de utilidad bipolar se normalizan para tener valores en el intervalo [0,100].

4.8.2 Funciones de utilidad basadas en generadores

Considere el método de construcción de CBSF U en la escala bipolar K = (-m, ..., -1, 0, 1, ..., m), m> 0. La función de utilidad bipolar sobre J puede ser construida además por (18) Y (20). Sea G un valor real positivo y g:{0, ..., m} → [0, G], es una función estrictamente creciente tal que g (0) = 0, g (m) = G. Esta función se llamará Un generador de la función W: $K \to [-G, G]$ definido por:

$$W(k) = g(k) \quad \text{for all } k \in \{0, \ldots, m\}, \qquad (22)$$

$$W(k) = -g(-k) \text{ for all } k \in \{-m, \ldots, -1\} \qquad (23)$$

Está claro que la función W construida por este método es la función de puntuación bipolar centrada. En el enfoque basado en modelos para modelar las preferencias del usuario podemos definir el generador g por una función paramétrica y usarlo para la definición de funciones paramétricas de utilidad bipolar sobre K o sobre J.

Abajo hay dos generadores paramétricos de funciones de utilidad bipolares inspiradas por la negación Sugeno paramétrica utilizada en la lógica difusa [Sugeno 1974], aquí (p> -1):

$$g_1(k) = \frac{Gk(1+p)}{m+pk}, \qquad k = 0, \ldots, m, \qquad (24)$$

$$g_2(k) = \frac{Gk}{m+pm-pk}, \quad k = 0, \ldots, m \qquad (25)$$

La figura 3 representa las formas de CBSF W en la escala bipolar centrada de 7 puntos K = (-3, -2, -1, 0, 1, 2, 3) obtenidos de los generadores (24) y (25) para diferentes valores de el parámetro p. Para p = 0 ambos generadores son lineales: $g_1(k) = g_2(k) = Gk/m$.

El generador para p positivo y el generador $g_2(k)$ para p negativo pueden utilizarse para modelar funciones de utilidad bipolares cuando cuanto más fuertes sean las categorías positivas o negativas de la escala menor será la diferencia entre las utilidades de las

categorías vecinas Las puntuaciones están situadas más cerca de los límites de la escala, véase la Fig. 3). Para dicha función de utilidad bipolar para la escala de 7 puntos

L = (absolutamente horrible, muy malo, malo, no bueno ni malo, bueno, muy bueno, excelente), no habrá grandes diferencias entre las utilidades de las categorías *"muy bueno"* y *"excelente"*. Por el contrario, el generador $g_1(k)$para p negativo y el generador $g_2(k)$ para p positivo pueden usarse para modelar funciones de utilidad bipolares de tal manera que cuanto más fuertes sean las categorías positivas o negativas de la escala mayor será la diferencia entre los valores de utilidad de Las categorías vecinas (las puntuaciones están situadas más cerca del centro de la escala). En la última escala la categoría "excelente" tendrá mucho más utilidad que la categoría "muy buena".

En el enfoque basado en modelos para el análisis de las opiniones de los usuarios se pueden ajustar los parámetros de las funciones de utilidad para obtener las mejores soluciones en la salida del modelo.

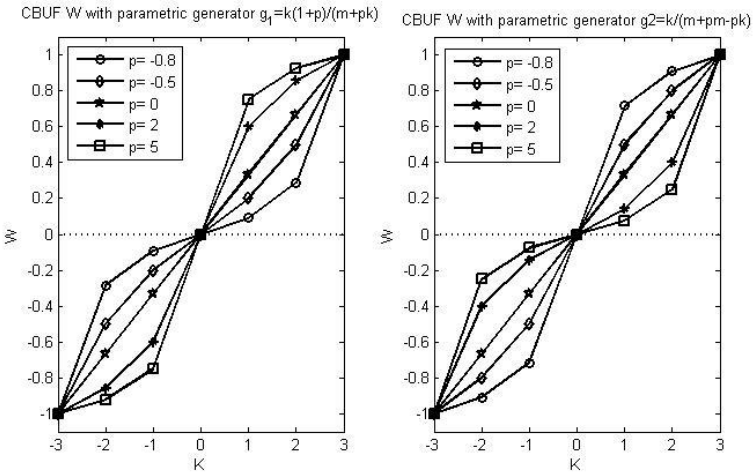

Figura 11. Ejemplos de CBSF U en escala bipolar centrada en 7 generados por generadores $g_1(k)$ (izquierda) y $g_2(k)$ (derecha) para diferentes valores del parámetro p.

4.8.2 *Selección paso a paso de los valores del generador*

A continuación se muestra el pseudocódigo del método general de construcción del generador de función de utilidad bipolar dado por (22) y (23). La función Determine depende del método específico, por ejemplo: selección aleatoria, basada en la optimización de algunos criterios, etc.

$g(0)=0$; $g(m) = G$;

para $k=1:m-1$

Determine $u \in (g(k-1), G)$

$g(k)=u$;

fin

5. METODOLOGIA DE EVALUACIÓN

En esta etapa se evaluara en primera instancia la eficiencia de la etapa del análisis sintáctico para lo cual se utilizara la medida de Attachment score. La cual según Vilar

(2013) se ha convertido en un estándar de evaluación para este análisis y su cometido es medir el porcentaje de palabras a las que se les ha asignado correctamente el padre en el árbol de dependencias. Se dispone de dos variantes de esta métrica, según se tenga en cuenta el tipo de dependencia asignada o no y se describe de la siguiente manera:

- LAS (Label Attachement Score): Métrica que mide el porcentaje de las palabras as laque tanto el padre como el tipo de dependencia fueron asignadas correctamente.
- UAS (Unlabeled Attachment Score): Métrica que solo tiene en cuenta que el padre de la relación de dependencia esté bien asignado.
- LAS2 (Label Acuracy Score): Mide solo que el tipo de dependencia haya sido correctamente anotados.

Para la evaluación del método de clasificación se utilizara la técnica de K-fold cross validation. Consiste en dividir todos los datos en 10 o k volúmenes. Se utilizara el primer volumen para la evaluación de los resultados del sistema, y los volúmenes de 2 a 10 para el

entrenamiento, después se elige el volumen dos para la evaluación y los otros nueve volúmenes para el entrenamiento y así sucesivamente (Sidorov, 2013).

El *baseline* se tomara a partir de la propuesta de Pang and Lee (2002) en la cual se realiza la detección de la polaridad a través de técnicas de aprendizaje automático como son naive bayes, máximum entropy classification y support vector machine (SVM). Para el caso de comparación se tomaran los resultados obtenidos por el clasificador de SVM los cuales son los de mejor precisión repórtados en el estudio.

Se toma como *baseline* ya que es de los trabajos iniciales que incorporan aprendizaje automático, utilizan un corpus para el entrenamiento y evaluación de criticas a películas de cine anotado manualmente por los autores, el cual se describió en el apartado de recursos léxicos de este documento, la similitud de los datos de entrenamiento y técnicas de clasificación son las que motivan su uso como *baseline.*

Otra de los factores por lo cual se selecciona este trabajo es su importancia en el estado del arte, el artículo de Pang and Lee (2002) cuenta con 2811 citas en google scholar, lo cual muestra que es un referente para cualquier trabajo de minería de opinión.

5.1 *Medidas de correlación en el conjunto de perfiles bipolares*

Demostremo que el coeficiente de correlación producto-momento de Pearson

$$corr(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{.....(26)}$$

A menudo utilizados en los sistemas de recomendación para medir la similitud entre los perfiles [***] puede ser engañoso si las opiniones se miden en las escalas bipolares. Supongamos que tenemos los siguientes perfiles de utilidad con las calificaciones de 10 ítems en la escala bipolar de 7 puntos J = {1, 2, 3, 4, 5, 6, 7} con la función de utilidad estándar $U(j) = j$ para todo j en j

x = (7,5,5,7,7,7,5,7,5,5),

y = (5,7,7,5,5,5,7,5,7,7),

z = (3,1,1,3,3,3,1,3,1,1).

Los perfiles x y y tienen sólo puntuaciones "positivas" (mayores que neutras $C = 4$) y la medida de asociación razonable debe dar la asociación positiva entre ellas: $A(x,y) > 0$ pero el coeficiente de correlación dado : $corr(x,y) = -1$. Los perfiles x y z tienen puntuaciones casi opuestas ("positivas" vs "negativas") y la medida de asociación razonable debe dar la asociación negativa entre ellas: $A(x,z) < 0$; pero el coeficiente de correlación dado $corr(x,z) = 1$.

Por lo tanto, necesitamos introducir medidas de correlación (asociación) que, como el coeficiente de correlación, pudieran medir asociaciones positivas y negativas entre perfiles de calificaciones en escalas bipolares pero que no tengan inconvenientes del coeficiente de correlación de Pearson como en el ejemplo anterior. Tales medidas se consideran a continuación. Se basan en los resultados generales discutidos en Batyrshin (2013), Batyrshin (2015).

Sea I una escala bipolar ($I = J$ or $I = K$) con la negación N y con el centro C. El vector $x = (x1, \ldots, xM)$, $xs \in I$, $s = 1, \ldots, M$, de elementos de la escala bipolar I se denominará un perfil de calificación. El vector $CX = (C, \ldots, C)$ de la longitud M se denominará perfil central del conjunto X de todos los perfiles de calificación de la longitud M. Denota la negación del perfil x como sigue: $NX(x) = (N(x1),\ldots, N(xM))$. Es claro que NX es una involución en X, es decir $NX\ (NX(x)) = x$, para todos los perfiles de X y CX es el punto fijo único de N, tal que $NX\ (CX) = CX$. Si en la escala bipolar I se define la función de utilidad bipolar U entonces el vector $U\ X\ (x) = (U\ (x\ 1), \ldots, U\ (x\ M))$ se llamará un perfil de utilidad del perfil de clasificación x.

Supongamos que un usuario evalúa 6 elementos en la escala de clasificación bipolar $J = \{1, 2, 3, 4, 5\}$ por el vector de valores $x = (3, 5, 2, 4, 1, 3)$. El perfil central del conjunto X de todos los perfiles de calificación de la longitud 6 estará dado por el vector CX = (3, 3, 3, 3, 3, 3). La negación del perfil x estará dada por el siguiente vector $N\ (x) = (3, 1, 4, 2, 5, 3)$. Si en la escala bipolar J se define la función de utilidad bipolar $U(J) = \{-10, -3, 0, 3, 10\}$, entonces el perfil de utilidad del perfil de clasificación x será dado por el vector: $U\ (X) = (0, 10, -3, 3, -10, 0)$. (Si no es confuso, aquí usamos las mismas letras N y U para designar las funciones definidas en J y en X).

Definiremos a continuación la medida de correlación $A_U(x,y)$ en función de los perfiles de utilidad $U(x)$ y $U(y)$ para los perfiles de calificación dados x e y con la misma longitud. En aplicaciones, cuando los perfiles de usuario tienen diferentes longitudes, los vectores x e y pueden contener únicamente calificaciones de elementos que tienen evaluaciones de ambos usuarios.

Definición 2. Sea X el conjunto de todos los perfiles de la longitud M con calificaciones de la escala bipolar I con la negación N y el centro C. Sea U una función de utilidad bipolar definida en I. Una medida de correlación (asociación) en el conjunto $V = X \setminus \{CX\}$ es una función A_U: $V \times V \rightarrow [-1, 1]$ satisfaciendo para todo $x, y \in V$ las siguientes propiedades:

$A_U(x,y) = A_U(y,x)$ (Simetría) ……(27)

$A_U(x,x) = 1$, (Reflexividad)……..(28)

$A_U(x,N(y)) = -A_U(x,y)$. (Relación inversa)……(29)

La medida de correlación A_U será llamada C-separable si satisface las siguientes propiedades:

$A_U(x,y) > 0$ si para todas las s=1,…, M esto es cumplido por $x_s, y_s > C$ or $x_s, y_s < C$,…….(30)

$A_U(x,y) < 0$ si para todas las s=1,…, M esto es cumplido por $y_s < C < x_s$ or $x_s < C < y_s$ …(31)

Las propiedades (27) - (29) utilizadas en la definición de la correlación (asociación) medida se han introducido en Batyr (2013, 2015) como la generalización de las propiedades del coeficiente de correlación de Pearson en cualquier conjunto con involución N. Como se mostró, El coeficiente de correlación puede ser extremadamente engañoso al medir las asociaciones entre los perfiles de clasificación en las escalas bipolares. Las propiedades (3) - (31) se introducen aquí para evitar los problemas con el coeficiente de correlación discutido al principio de esta sección. Véase también la propiedad de separabilidad C en [0,1] considerada en Batyr (2015).

De (28) y (29):

$A_U(x,N(x)) = -1$……(32)

Definición 2 se puede extender para el conjunto V en todos los perfiles X que reemplaza la propiedad (28) por el siguiente:

$A_U(x,x) = 1$, if $x \neq C_X$ ……(33)

En este caso se cumple para todo x en X:

$A_U(x,C_X) = A_U(C_X,x) = 0$……(34)

Consideremos los métodos de construcción de medidas de asociación de medidas de correlación en el conjunto de perfiles con valores bipolares basados en métodos discutidos en Batyrshin (2013).

Proposición 2. Sea I una escala bipolar ($I = J$ *o* $I = K$) con el centro C, X un conjunto de perfiles $x = (x_1,\ldots, x_M)$ de la longitud M, $x_s \in I$, $s = 1, \ldots, M$, con el perfil central $C_X = (C, \ldots, C)$, y U es una función de puntuación bipolar en I, entonces la siguiente función es medida de correlación C separable en $X \setminus \{CX\}$:

$$A_U(x,y) = \frac{1}{2^t}\sum_{s=1}^{M}[|F(x_s) + F(y_s)|^t - |F(x_s) - F(y_s)|^t] \quad \ldots\ldots (35)$$

where

$$F(x_s) = \frac{U(x_s) - U(C)}{\sqrt[t]{\sum_{s=1}^{M}|U(x_s) - U(C)|^t}} \quad \ldots\ldots (36)$$

y $t \geq 1$.

Se puede comprobar fácilmente que se cumplen las propiedades (27) - (31) para (35). Si la función de puntuación bipolar U en (36) está centrada, es decir $U(C) = 0$, entonces (36) se simplifica como sigue:

$$F(x_s) = \frac{U(x_s)}{\sqrt[t]{\sum_{s=1}^{M}|U(x_s)|^t}} \quad \ldots\ldots (37)$$

La Proposición 2 implica lo siguiente

Corolario 1. Si en las condiciones de la Proposición 2, if $t = 2$ in (35), (36), entonces la siguiente función es medida de correlación C-separable en X \ {CX}:

$$A_U(x,y)=\frac{\sum_{s=1}^{M}(U(x_s)-U(C))(U(y_s)-U(C))}{\sqrt{\sum_{s=1}^{M}(U(x_s)-U(C))^2}\sqrt{\sum_{s=1}^{M}(U(y_s)-U(C))^2}} \quad \text{.....(38)}$$

Si la función de puntuación bipolar U en (38) está centrada, entonces:

$$A_U(x,y)=\cos(U(x),U(y))=\frac{\sum_{s=1}^{M}U(x_s)U(y_s)}{\sqrt{\sum_{s=1}^{M}U(x_s)^2}\sqrt{\sum_{s=1}^{M}U(y_s)^2}} \quad \text{.....(39)}$$

Debido a que la fórmula (37) y (37) requieren un menor número de operaciones que las fórmulas (36) y (38) en el cálculo de la correlación $A_U(x,y)$ entre el gran número de pares de perfiles (x, y) se recomienda reemplazar la función de utilidad bipolar U Definida en la escala bipolar I por la función de utilidad centrada U-$U(C)$ y luego calcular las correlaciones entre perfiles correspondientes por (35),(37) o (39).

Calculemos la correlación $A_U(x,y)$ entre los perfiles de las clasificaciones x = (7,5,5,7,7,7,5,7,5,5), y = (5,7,7,5,5 , 5,7,5,7,7), z = (3,1,1,3,3,3,1,3,1,1) a partir de la escala bipolar $J = \{1, 2, 3, 4, 5 , 6, 7\}$ considerados al principio de esta sección. Para la función de utilidad estándar $U(J) = J$, obtenemos $U(x) = x$, $U(y) = y$, $U(z) = z$. Podemos calcular la correlación entre tres perfiles por (38). Sin embargo, como hemos señalado anteriormente, es más eficiente reemplazar la función de utilidad bipolar U por la función de utilidad centrada $U_C(J)= U(J) - U(C)= J - 4 = K = \{-3, -2, -1, 0, 1, 2, 3\}$ y utilizar esta función de utilidad centrada en (39). Obtenemos: $A_U(x, y) = 0.6$, $A_U(y, z) = -1$, $A_U(x, z) = -0.6$. Estos valores corresponden a nuestras proposiciones $A(x, y) > 0$ and $A(x, z) < 0$ consideradas al comienzo de esta sección, y la nueva medida de correlación (38) no presenta los inconvenientes del coeficiente de correlación de Pearson. Tenga en cuenta que tenemos $z = N(y)$, y por esta razón, según la propiedad (29) de la medida de correlación obtenemos: $A_U(x, z) = A_U(x, N(y)) = -A_U(x, y) = -0.6$; De acuerdo con la propiedad (32), obtenemos $A_U(y, z) = A_U(y, N(y)) = -1$.

La medida de correlación (38) obtenida aquí como caso particular de (35) generaliza el coeficiente de correlación restringido considerado en [32] (véase la fórmula (5)) usando en (38) la función de utilidad estándar de 7 puntos $U = J$ *con centro* $C = 4$ y $U(C) = 4$.

Considere (38) en la forma (39) cuando la función de utilidad U es reemplazada por su forma centrada $U - U(C)$. Como se puede ver, la fórmula (39) es una función no lineal sensible a la presencia de valores de utilidad respectivamente altos para los mismos elementos en ambos perfiles x e y.

Considere un ejemplo para la escala de 7 puntos: $K = \{-3, -2, -1, 0, 1, 2, 3\}$. Supongamos que uno utiliza la función de utilidad estándar $U(K) = K$, y dos usuarios tienen los siguientes perfiles de clasificación de cuatro ítems en la escala K: $x = \{1, 1, 1, 1\}$, $y = \{-1, -1, -1, -1\}$. Debido a $U(K) = K$, los perfiles de utilidad correspondientes tendrán los mismos valores $U(x) = x$, $U(y) = y$. Puesto que los perfiles son opuestos: $N(x) = y$, *la correlación (39) entre ellos tiene el valor* $A(x,y) = -1$. Supongamos que un nuevo elemento tiene la calificación 2 de ambos usuarios. Entonces obtenemos los nuevos perfiles: $x^* = \{1, 1, 1, 1, 2\}$, $y^* = \{-1, -1, -1, -1, 2\}$ Como se puede ver, la adición del mismo valor 2 a ambas clasificaciones cambia drásticamente el valor de correlación de -1 a 0. Esta situación se puede evitar si usamos la función de utilidad bipolar no lineal. Sea $U(1) = 1$ and $U(2) = 1.5$. Entonces obtenemos los siguientes perfiles de utilidad: $x^* = \{1, 1, 1, 1, 1.5\}$, $y^* = \{-1, -1, -1, -1, 1.5\}$ con correlación $A(x^*, y^*) = -0.28$. Como se puede ver, el cambio del valor de correlación de -1 a -0,28 no es tan drástico para la función de utilidad bipolar no lineal considerada como para la función de utilidad lineal estándar.

Consideremos otro ejemplo de perfiles con calificaciones de la escala de calificación de 7 puntos K. Supongamos que usamos la función de utilidad estándar $U(K) = K$ *y dos usuarios tienen perfiles de calificación iguales de cinco elementos:* $x = \{1, 1, 1, 1, 2\}$, $y = \{1, 1, 1, 1, 2\}$. Tenemos $A(x, y) = 1$. Supongamos que para el sexto punto ambos usuarios tienen las calificaciones opuestas 3 y -3. Para la función de utilidad estándar, la correlación entre los nuevos perfiles de utilidad $x^* = \{1, 1, 1, 1, 2, 3\}$ y $y^* = \{1, 1, 1, 1, 2, -3\}$ cambia drásticamente, Desde el valor $A(x, y) = 1$ hasta el valor $A(x^*, y^*) = -0.059$. Cambiemos la función de utilidad estándar por $U(K) = \{-2, -1.5, -1, 0, 1, 1.5, 2\}$. Para esta función de utilidad bipolar no lineal, la correlación entre perfiles cambia de $A(x, y) = 1$ to $A(x^*, y^*) = 0.22$, que no es tan drástica como para la función de utilidad estándar.

En los dos ejemplos considerados, se utilizó el método de construcción de funciones de utilidad bipolar considerado en la sección 6.3, que define la parte positiva de la función

de utilidad centrada y simétricamente la asigna a la parte negativa de la escala con 0 en el centro de la escala . Del mismo modo, hemos definido secuencialmente $U(1) = 1$, $U(2) = 1.5$, and $U(3) = 2$. Otro método puede basarse en un generador paramétrico como (24) o (25). Por ejemplo, utilizando el generador (24) con el parámetro $p = 1$, podemos obtener resultados similares: para la función de utilidad bipolar no lineal $U(K) = \{-2, -1.6, -1, 0, 1, 1.6, 2\}$, La correlación entre los perfiles x^* y y^* *tiene el valor* $A(x^*, y^*) = 0.24$.

En ambos ejemplos, se disminuyeron los valores de utilidad absoluta de las categorías cerca de los polos para evitar el cambio drástico del valor de correlación cuando ambos usuarios usan clasificaciones casi polares para los mismos ítems. En general, una función de utilidad paramétrica puede ser heurísticamente ajustada para obtener soluciones intuitivamente validadas u optimizadas por algún método de aprendizaje máquina, para los sistemas de recomendación o toma de decisiones utilizando escalas de clasificación bipolares para obtener soluciones con mejor rendimiento.

6. Discusiones y Conclusiones

Las escalas de calificación aplicadas en diferentes áreas de aplicación suelen tener estructura bipolar, pero esta bipolaridad entre categorías opuestas de las escalas a menudo no fue explícita o explotada formalmente. El presente trabajo introduce explícitamente la propiedad de la bipolaridad en la definición de la estructura general de las escalas bipolares verbales, en la definición formal de la escala bipolar como el conjunto ordenado linealmente de índices con operación de negación y en la definición de la función de puntuación bipolar en la conservación de la escala bipolar La simetría de esta escala. En la descripción de la estructura general de las escalas bipolares en la Sección 3, basamos en el documento (Batyrshin,1990).

La idea de la definición formal de la escala bipolar en el conjunto de índices se basó parcialmente en el artículo de Herrera(2000) donde las categorías lingüísticas de la escala se presentan por conjuntos difusos. En la sección 4 consideramos juntos dos conjuntos de índices mutuamente relacionados donde $J = \{1,\ldots, 2m+1\}$ es más tradicional y $K= \{-m, \ldots, m\}$ es más "natural" para la representación de escalas bipolares con opuestos Categorías.

La mayoría de las escalas bipolares observadas en el capítulo 3 pueden representarse como escalas bipolares con estructura general consideradas en ese mismo capítulo o formalmente como escalas bipolares consideradas en el capítulo 4. El concepto de puntuación bipolar o función de utilidad definida en la escala bipolar es la mejor aportación de este trabajo. Las funciones de puntuación bipolar incluyen la puntuación tradicional de escalas de puntuación de n puntos por números *1, ..., n* como caso particular y tales funciones de puntuación se denominan funciones de puntuación bipolares estándar. Pero parece más interesante en lugar de las funciones de puntuación estándar o en lugar de las funciones de utilidad lineal considerar funciones de utilidad no lineales.

Estas funciones de utilidad pueden ser dadas como funciones paramétricas y ajustadas por algún procedimiento de aprendizaje mecánico para obtener resultados óptimos o buenos en la salida del sistema de recomendación o toma de decisiones usando estas escalas bipolares. Las funciones de puntuación bipolar no lineal pueden ser útiles en el modelado de las calificaciones de los usuarios o en la utilidad de modelado o la importancia de las categorías en las escalas bipolares. Las razones para utilizar las funciones de utilidad no lineal bipolar en la medida de correlación introducida en el documento se tratan en el capítulo 5. Los resultados sobre las medidas de asociación consideradas en el capítulo 5 se basan en los documentos de Batyrshin (2013) y Batyrshin (2015).

Se utilizaron los términos de medida de asociación y medida de correlación como intercambiables. Se extendio la propiedad de *C-separabilidad* de la medida de asociación en [0,1] considerado por Batyrshin (2015) en el conjunto de perfiles de utilidad. La fórmula general para la medida de asociación en los perfiles de utilidad bipolar se basa en la distancia de Minkowski y en los resultados generales considerados por Batyrshin (2013) para las series de tiempo. Las fórmulas de Thurstone (1928) y Weijters (2010) para funciones de utilidad bipolar centradas son específicas para las medidas de asociación separable en C en perfiles de utilidad bipolar. La fórmula de Weijters (2010)generaliza la medida de correlación considerada en Shardanand(1995) sin funciones de utilidad. En nuestro trabajo futuro planeamos aplicar los resultados del trabajo en el filtrado colaborativo y en el análisis de las calificaciones humanas.

7. REFERENCIAS BIBLIOGRÁFICAS

1. Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 26(3), 12.

2. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), pp. 734-749.

3. Assaf D, Neuman Y, Cohen Y, Argamon S, Howard N, et al.. (2013) Why "dark thoughts" aren't really dark: A novel algorithm for metaphor identification. To appear in the Proceedings of the IEEE Symposium Series on Computational Intelligence, Singapore, April 2013.

4. Atanassov, K. T. (1986). Intuitionistic fuzzy sets. Fuzzy sets and Systems, 20 (1), 87-96.

5. Aue, A., & Gamon, M. (2005, September). Customizing sentiment classifiers to new domains: A case study. In Proceedings of recent advances in natural language processing (RANLP) (Vol. 1, No. 3.1, pp. 2-1).

6. Batyrshin, I.Z. (1990). On the structure of verbal scales. In: Proceedings of the Second All-Union Conference on Artificial intelligence. Minsk, 1990, vol. 1, pp. 37-40.

7. Batyrshin I. (2011). Uncertainties with memory in construction of strict monotonic t-norms and t-conorms for finite ordinal scales: basic definitions and applications. Applied and Computational Mathematics, vol. 10, 3, 2011, pp. 498-513.

8. Batyrshin, I. (2013). Constructing time series shape association measures: Minkowski distance and data standardization. 2013, https://arxiv.org/abs/1311.1958v3.

9. Batyrshin, I. Z. (2015). On definition and construction of association measures. Journal of Intelligent & Fuzzy Systems, 29(6), 2015, pp. 2319-2326.

10. Breese, J. S., Heckerman, D., & Kadie, C. (1998, July). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (pp. 43-52). Morgan Kaufmann Publishers Inc.

11. Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2009). EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. In DMIN (pp. 491-497).

12. Brill, E. (1994). Some advances in transformation-based part of speech tagging.arXiv preprint cmp-lg/9406010.

13. Buchholz, S., & Marsi, E. (2006, June). CoNLL-X shared task on multilingual dependency parsing. In Proceedings of the Tenth Conference on Computational Natural Language Learning (pp. 149-164). Association for Computational Linguistics.

14. Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In Proceedings of the twenty-eighth AAAI conference on artificial intelligence (pp. 1515-1521). AAAI Press.

15. Carbonell, J. G. (1979). Subjective Understanding: Computer Models of Belief Systems (No. RR-150). YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE.

16. C. Strapparava and R. Mihalcea. SemEval-2007 Task 14: Affective Text. In Proceedings of the thInternational Workshopon Semantic Evaluations (SemEval 2007), Prague, Czeck Republic, pages 70–74, 2007.

17. Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. Applied Psychological Measurement, 18(3), 205-215.

18. Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29.

19. Chomsky, N. (2002). Syntactic structures. Walter de Gruyter.

20. Cruz, F. L., Troyano, J. A., Enriquez, F., & Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol. Procesamiento de Lenguaje Natural, 41.

21. Das, S., & Chen, M. (2001, August). Yahoo! for Amazon: Sentiment parsing from small talk on the Web. In EFA 2001 Barcelona Meetings.

22. Dave, K., S. Lawrence, and D. M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

23. Dubois, D., & Prade, H. (2006). Bipolar representations in reasoning, knowledge extraction and decision processes. In International Conference on Rough Sets and Current Trends in Computing, pp. 15-26. Springer Berlin Heidelberg.

24. E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions, *Proceeding of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-2003), pages 105-112, 2003.

25. Fernandez Anta, A. (2012). Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report.

26. Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. Personality and Individual Differences, 40(5), 873-884.

27. Goatly , A. (1997). The Language of Metaphors. London: Routledge.

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

29. Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. information retrieval, 4(2), 133-151.

30. Grabisch, M. (2006). Aggregation on bipolar scales. In Theory and applications of relational structures as knowledge instruments II (pp. 355-371). Springer Berlin Heidelberg.

31. Gunderman, R. B., & Chan, S. (2013). The 13-point Likert scale: a breakthrough in educational assessment. Academic radiology, 20(11), 2013, pp. 1466-1467.

32. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

33. Haro, S. N. G., & Gelbukh, A. (2007). Investigaciones en análisis sintáctico para el español. Instituto Politécnico Nacional, Dirección de Publicaciones.

34. Hashimoto, Chikara, and Daisuke Kawahara. "Compilation of an idiom example database for supervised idiom identification. "Language resources and evaluation 43.4 (2009): 355-384.

35. Herrera, F., & Herrera-Viedma, E. (2000). Linguistic decision analysis: steps for solving decision problems under linguistic information. Fuzzy Sets and systems, 115(1), 67-82.

36. Hjermstad, M. J., Fayers, P. M., Haugen, et al (2011). Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: a systematic literature review. Journal of Pain and Symptom Management, 41(6), 1073-1093.

37. J. Wiebe and R. Mihalcea. Word sense and subjectivity. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL,pages 1065–1072, 2006.

38. Jang, J. S. R., Sun, C. T., & Mizutani, E. (1997). Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Prentice Hall.

39. Jia, L., Yu, C., & Meng, W. (2009, November). The effect of negation on sentiment analysis and retrieval effectiveness. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1827-1830). ACM.

40. Juniper, E. F., Guyatt, G. H., Ferrie, P. J., & King, D. R. (1999). Development and validation of a questionnaire to measure asthma control. European Respiratory Journal, 14(4), 902-907.

41. Jhonson, P. (2005). Voting systems. A textbook-style overview of voting methods and their mathematical properties.

42. Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2), 110-125.

43. Kim, S-M. and E. Hovy. 2004. Determining the sentiment of opinions. In Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics.

44. Klepousniotou E, Titone D, Romero C (2008) Making sense of word senses: The comprehension of polysemy depends of sense overlap. J Exp Psychol Learn 34: 1534–1543. doi: 10.1037/a0013012.

45. Krishnakumaran, S., Zhu, X. (2007). Hunting elusive methapors using lexical resources. In: Proceedings of the Workshop on Computational Approaches to Figurative Language (FigLanguages ‘07). Stroudsburg, PA: Association for Computational Linguistics. 13-20.

46. L. Polanyi and A. Zaenen. Contextual valence shifters. Computing Attitude and Affect in Text: Theory and Applications, pages 1–9, 2004.

47. L. V. Lita, A. H. Schlaikjer, W. Hong, and E. Nyberg, “Qualitative dimensions in question answering: Extending the definitional QA task,” in Proceedings of AAAI, pp. 1616–1617, 2005.

48. Lakoff G, Johnson M (1980) Metaphors we live by. Chicago: University of Chicago Press.

49. Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology. Vol 22, No. 140, 1932, pp. 55. New York.

50. Liu, J., & Seneff, S. (2009). Review sentiment scoring via a parse-and-paraphrase paradigm. In Proceed. 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 1 (pp. 161-169). Association for Computational Linguistics.

51. Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2, 627-666.

52. Malvar-Fernández, P., & Pichel-Campos, J. R. (2011). Generación semiautomática de recursos de Opinion Mining para el gallego a partir del portugués y el español. In Proceedings of the workshop on Iberian cross-language natural language processing tasks (ICL 2011)(pp. 59-63).

53. Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

54. Martínez Cámara, E., Martín Valdivia, M. T., Perea Ortega, J. M., & Ureña López, L. A. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español.

55. Martín-Valdivia, M. T., Martínez-Cámara, E., Perea-Ortega, J. M., & Alfonso Ureña-López, L. (2012). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications.

56. Moreno-Ortiz, A., & Hernández, C. P. (2012). Lexicon-based sentiment analysis of Twitter messages in Spanish. Procesamiento del Lenguaje Natural, 50, 93-100.

57. Morsy, S. A., & Rafea, A. (2012). Improving document-level sentiment classification using contextual valence shifters. In Natural Language Processing and Information Systems (pp. 253-258). Springer Berlin Heidelberg.

58. Mosier, C.I. (1941). A psychometric study of meaning. The Journal of Social Psychology, 13(1), 123-140.

59. Mullen, T., & Collier, N. (2004, July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp. 412-418).

60. Neuman Y, Assaf D, Cohen Y, Last M, Argamon S, et al. (2013) Metaphor Identification in Large Texts Corpora. PLoS ONE 8(4): e62343. doi:10.1371/journal.pone.0062343.

61. Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages. 79–86.

62. Ortiz, A. M., Castillo, F. P., & García, R. H. (2010). Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. Procesamiento del lenguaje natural, 45, 31-39.

63. Osgood, C.E. (1952). The nature and measurement of meaning. Psychological bulletin, 49(3), 197-237.

64. Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).

65. Pan, L. (2012). Sentiment Analysis in Chinese (Doctoral dissertation, Brandeis University).

66. Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 115-124). Association for Computational Linguistics.

67. Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

68. Petrenko V.F. (1988). Psychosemantics of consciousness. Moscow, Ripol Klassik. (In Russian: Петренко, В. Ф. (1988). Психосемантика сознания. Рипол Классик).

69. Perea-Ortega, J. M., & Balahur, A. (2014). Experiments on feature replacements for polarity classification of spanish tweets. In Proceedings of the TASS workshop at SEPLN.

70. Polanyi, L. and A. Zaenen. 2006. Contextual valence shifters. In W. Bruce Croft, James Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*. Springer Netherlands, pages 1–10.

71. Pfanzagl, J. (1971). Theory of measurement. Physica. Physica-Verlag Heidelberg.

72. Poria S, Gelbukh A, Cambria E, Hussain A & Huang G (2014). EmoSenticSpace: A novel framework for affective commonsense reasoning, Knowledge-Based Systems, 69, pp. 108-123.

73. Pospelov D.A. (1989). Models of Reasoning. Essay in the Analysis of Mental Acts. Radio y Svyaz. Moscow. (In Russian: Поспелов Д.А. Моделирование рассуждений. Опыт анализа мыслительных актов , М.: Радио и связь, 1989, -184с.)

74. Ricci, F., Rokach, L., Shapira, B., Kantor P.B. (2011) Recommender Systems Handbook, Springer US.

75. Roberts, F. S. (1985). Measurement theory. Cambridge University Press.

76. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F. & Gordon, J. (2013). Empirical study of machine learning based approach for opinion mining in tweets. In Advances in Artificial Intelligence (pp. 1-14). Springer Berlin Heidelberg.

77. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2013). Syntactic dependency-based n-grams as classification features. In dvances in Computational Intelligence (pp. 1-11). Springer Berlin Heidelberg.

78. Sugeno M. (1974) Thery of Fuzzy Integrals and its Applications. Dissertation. Tokio Institute of Technology.

79. Schafer, J. H. J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. The adaptive web, 291-324.

80. Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth". In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 210-217). ACM Press/Addison-Wesley Publishing Co.

81. Shaikh, M. A. M., Prendinger, H., & Ishizuka, M. (2008). Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. Applied Artificial Intelligence, 22(6), 558-601.

82. Shutova, E., Teufel, S., Korhonen, A. (2012). Statical metaphor processing. Computational Linguistic. In press. Doi: 10.1162/COLI_a_00124.

83. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. EMNLP.

84. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

85. Tarassov V.B. (2001). Analysis and modeling of NOT-factors on polar scales. In: Integrated Models and Soft Computing in Artificial Intelligence, Moscow, Fismatlit, pp. 65-71. (In Russian: Тарасов В.Б. Анализ и моделирование НЕ-факторов на полярных шкалах// Интегрированные модели и мягкие вычисления в искусственном интеллекте. − М.: Наука. Физматлит, 2001. − С.65-71).

86. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544-2558.

87. Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 33(4), 529-554.

88. Turnet, P. , Neuman, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In: Proceedings of the 2011 Conference on Empirical

Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27-31:680-690.

89. Turney, P. 2002. Thumbs up or thumbs down: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics, pages 417–424.

90. V. Rentoumi, V. Karkaletsis, G. Vouros, and A. Mozer. Sentiment analysis exploring metaphorical and idiomatic senses: A word sense disambiguation approach. Procceding of International Workshop on Computational Aspects of Affectual and Emotional Interaction (CAFFEi 2008), 2008.

91. Vilares Calvo, D., Alonso Pardo, M. Á., & Gómez Rodríguez, C. (2013). Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias.

92. Wan, X. (2009, August). Co-training for cross-lingual sentiment classification. InProceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 235-243). Association for Computational Linguistics.

93. Wang, H., Liu, L., Song, W., & Lu, J. (2014). Feature-based Sentiment Analysis Approach for Product Reviews. Journal of Software, 9(2), 274-279.

94. Wang, M., & Shi, H. (2010, December). Research on sentiment analysis technology and polarity computation of sentiment words. In Progress in Informatics and Computing (PIC), 2010 IEEE International Conference on (Vol. 1, pp. 331-334). IEEE.

95. Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. International Journal of Research in Marketing, 27(3), 236-247.

96. Wiebe, J. M. (1994). Tracking point of view in narrative. Computational Linguistics, 20(2), 233-287.

97. Wiebe, J. M., T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational Linguistics*, 30:277–308.

98. Wilks, Y., & Bien, J. (1983). Beliefs, points of view, and multiple environments.Cognitive Science, 7(2), 95-119.

99. Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 347-354). Association for Computational Linguistics.

100. Wilson, T., J. M. Wiebe, and P. Hoffmann. 2009. Recognized contextual polarity: An exploration of features for phrase-level sentiment analisys. Computation Linguistics, 35 (3): 399- 433.

101. X. Jin, Y. Li, T. Mah, and J. Tong, “Sensitive webpage classification for content advertising,” in Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising, 2007.

102. Xu, Z. (2012). Linguistic decision making. Springer Berlin Heidelber.

103. Zadeh, L. A. The concept of a linguistic variable and its application to approximate reasoning.—I. Information sciences, 8(3), 1975, pp. 199-249; —II. Information sciences, 8(4), 1975, pp. 301-357. —III. Information sciences, 9(1), 1975, pp. 43-80.