Pontificia Universidad Católica de Valparaíso

Facultad de Ingeniería

Escuela de Ingeniería Informática

Clasificación automática de Tweets utilizando K-NN y K-Means como algoritmos de clasificación automática, aplicando TF-IDF y TF-RFL para las ponderaciones

FELIPE ALBERTO CIFUENTES RAMOS

Profesor Guía: **Rodrigo Alfaro Arancibia**

Profesor Co-referente: **Wenceslao Palma Muñoz**

Carrera: **Ingeniería Civil en Informática**

Agosto de 2016

Agradecimientos: Agradezco a mí hermano Héctor, por ser un pilar de apoyo en los momentos que más lo necesitaba, a mi madre Julia por siempre creer que yo podía y a mi padre Hernán por el apoyo que siempre me ha brindado. Agradezco también a todos mis amigos tanto los que conocí a través del camino universitario como a mis viejos amigos, ya que sin ellos no hubiese podido llegar tan lejos.

Índice

Índice i
Resumen iv
Lista de Figuras v
Lista de tablas vi
1 Introducción 1
2 Objetivos 3
2.1. Objetivo General 3
2.2. Objetivos Específicos 3
3 Metodología de Estudio 4
4 Problemática 5
5 Análisis de Sentimiento 6
5.1. Detección de Polaridad 6
5.2. Análisis del sentimiento basado en características 7
6 Clasificación automática de Textos 8
7 Representación de los datos 9
7.1. Pre- Procesamiento de los Datos 9
7.2. Caracterización de los datos 10
7.2.1. N-Gramas 10
7.2.2. Ponderado Booleano 11
7.2.3. Ponderado de frecuencia de término (TF) 11
7.2.4. Ponderado por (TF-IDF) 11
7.2.5. Ponderado por TF-RFL 11
7.2.6. Ponderado por TF-RRFL 12
7.3. Aprendizaje computacional 12
7.4. Taxonomía de sistemas de aprendizaje 12
7.4.1. Aprendizaje Supervisado 12
7.4.2. Aprendizaje No supervisado 13
7.4.3. Aprendizaje Semi- Supervisado 13
7.5. Tipos de aprendizaje 13
8 Aprendizaje Inductivo Supervisado 15
8.1. KNN o K-vecinos más cercanos 15
8.2. Elección de la variable K 15
8.3. Algoritmo KNN 17

8.4. Ventajas y Desventajas del algoritmo K-NN 17
9 Clustering 18
9.1. K-Means 19
9.2. Algoritmo K-Means 19
9.3. Elección de los K objetos y sus centroides 20
10 Distancias estadísticas 21
10.1. Distancia Euclidiana 21
10.2. Distancia Euclidiana Normalizada 21
11 Análisis de resultados 23
12 Corpus 25
12.1. Contexto de Experimentación 25
12.1.1. WEKA 25
12.2. Pasos de la experimentación 26
12.2.1. Experimentación con primer Corpus (Tweets): 26
12.2.2. Experimentación con segundo Corpus (Reuters): 26
13 Resultados obtenidos (Twitter) 27
13.1. Algoritmo K-NN 27
13.1.1. K-NN Sin Lematizar 27
13.1.2. K-NN Lematizado 28
13.1.3. Comparación de resultados según Lematización (K-NN) 29
13.2. Algoritmo K-Means 29
13.3. Conclusión primera experimentación 30
14 Resultados obtenidos (Reuters) 31
14.1. Algoritmo K-NN 31
14.2. Conclusión segunda experimentación 33
15 Trabajo Futuro 35
16 Conclusión 36
17 Bibliografía 38
Anexos 41
Anexo A: Resultados K-NN 41
Anexo A.1: K-NN Sin lematizar 41
Anexo A.2 K-NN Lematizado 42
Anexo B: K-Means 44
Anexo C: Tablas datos K-NN Sin Lematizar 44

Anexo D: Tablas datos K-NN Lematizado 47
Anexo E: REUTERS 50
Anexo E.1: Gráficos comparativos Reuters 51

Resumen

La presente tesis realiza el análisis y evaluación del desempeño que entrega la clasificación automática de texto utilizando los algoritmos K-NN y K-Means al ser aplicado a la minería de opinión. Estos algoritmos se utilizaron bajo diversos contextos lingüísticos (Lematización, Bi-gramas, Trigramas). A diferencia de estudios similares, se incorporó la ponderación TF-IDF y la TF-RFL como un medio de optimizar resultados esperados.

Los algoritmos fueron implementados y comparados para mostrar la relación y diferencia entre ellos al ser aplicados a un conjunto de Tweets, los cuales cuentan con las opiniones generadas por los usuarios de la plataforma Twitter en relación a una empresa de marketing, Falabella.

Palabras claves: Twitter, Análisis de Sentimiento, Caracterización de datos, TF-IDF, TF-RFL, Aprendizaje Computacional, K-NN, K-Means.

Abstract

The present thesis performs the analysis and evaluation of the performance that gives the automatic classification of text using the algorithms K-NN and K-Means when being applied to the opinion mining. These algorithms were used under different linguistic contexts (Lemmatization, Bi-grams, and Trigrams). Unlike similar studies, TF-IDF weighting and TF-RFL were incorporated as a means of optimizing expected results.

The algorithms were implemented and compared to show the relationship and difference between them when applied to a set of Tweets, which have the opinions generated by users of the Twitter platform in relation to a marketing company, Falabella.

Key words: Twitter, Sentiment analysis, Characterization of data, TF-IDF, TF-RFL, Computational Learning, K-NN, K-Means.

Lista de Figuras

Figura 7.1 Conjunto de Datos 9
Figura 7.2 Representación Unigrama 10
Figura 7.3 Representación Bi-grama 10
Figura 7.4 Representación Tri-grama 10
Figura 8.1 Representación de K 15
Figura 8.2 K grande 16
Figura 13.1 Promedio K-NN sin Lematizar (Vecindad) 27
Figura 13.2 Promedio según vecindad K-NN (Lematizado) 28
Figura 13.3 K-NN promedio Lematizado vs Sin Lematizar 29
Figura 13.4 K-Means comparativo 30
Figura 14.1 Promedios Ponderación y porcentajes 31
Figura 14.2 Promedios Ponderación y constante K 32
Figura 14.3 Promedio según K 33
Figura 0.1 70% Sin Lematizar 41
Figura 0.2 60% Sin Lematizar 41
Figura 0.3 50% Sin Lematizar 42
Figura 0.4 70% Lematizado 42
Figura 0.5 60% Lematizado 43
Figura 0.6 50% Lematizado 43
Figura 0.7 Promedio Reuters según K 51
Figura 0.8 Promedio Reuters según Ponderación 52
Figura 0.9 Promedio Reuters según Porcentaje 52

Lista de tablas

Tabla 0.1 Datos resumen K-Means 44
Tabla 0.2 Original: IDF 44
Tabla 0.3 Original: TF-RFL 44
Tabla 0.4 Bigrama: IDF 45
Tabla 0.5 Bigrama: TF-RFL 45
Tabla 0.6 Trigrama: IDF 45
Tabla 0.7 Trigrama: TF-RFL 45
Tabla 0.8 Desv-TF-RFL-Original 46
Tabla 0.9 Desv-TF-RFL-Bigrama 46
Tabla 0.10 Desv- TF-RFL- Trigrama 46
Tabla 0.11 95%-TF-RFL- Trigrama 46
Tabla 0.12 90% -TF-RFL- Trigrama 47
Tabla 0.13 Original: IDF 47
Tabla 0.14 Original: TF-RFL 47
Tabla 0.15 Bigrama: IDF 47
Tabla 0.16 Bigrama: TF-RFL 48
Tabla 0.17 Trigrama: IDF 48
Tabla 0.18 Trigrama: TF-RFL 48
Tabla 0.19 Desv-TF-RFL-Original 48
Tabla 0.20 Desv-TF-RFL-Bigrama 49
Tabla 0.21 Desv- TF-RFL- Trigrama 49
Tabla 0.22 95%-TF-RFL- Trigrama 49
Tabla 0.23 90% -TF-RFL- Trigrama 49
Tabla 0.24 Reuters 50% 50
Tabla 0.25 Reuters 60% 50
Tabla 0.26 Reuters 70% 51

1 Introducción

La red de internet es cada día más grande y los avances tecnológicos permiten hoy en día que se pueda acceder a esta red desde cualquier parte, en el momento que se necesite y para lo que se estime conveniente; por ende, este servicio ha pasado a ser indispensable en el que hacer de las personas en la actualidad, desde un niño que navega para jugar, hasta un adulto que lo utiliza para trabajar. Las funcionalidades que nos puede entregar la Internet son muy variadas permitiendo hacer casi cualquier cosa con solo buscar lo que se necesite.

De este modo, el número de internautas ha pasado de 1.000 millones de personas en el 2005 a 3.200 millones el año 2015, equivalente al 40% de la población mundial, llegando al 75% de la población en Europa. China el país más poblado del planeta supero los 420 millones de navegantes. Esto sumado a la globalización de los servicios e información disponibles a solo un "clic" de distancia, permite hacer una idea de la cantidad de usos que se le pueden dar. [1]

Uno de los servicios que más fuerza a adquirido en los últimos años es el de los microbloging, siendo Twitter uno de sus exponentes más importantes. Desde su lanzamiento el año 2007, ha ganado una gran popularidad estimándose unos 316 millones de usuarios mensualmente, generando 500 millones de tweets al día. Esta red permite enviar mensajes de corta longitud, llamados tweets, siendo publicadas en las páginas principales de los usuarios. Estos tweets generan un impacto en los seguidores ya sea positiva o negativa del tema en cuestión viralizando la publicación en torno a una etiqueta o hashtag. [2]

En el portal las personas están en total libertad de opinar, informar o preguntar acerca de algún tema que les compete. Debido a esto, la gran cantidad de información a ser extraída o inferida de estas publicaciones puede ser muy beneficiosa para algún mercado u organización, esto hace que este ámbito de la minería de datos, llamada minería de opinión se encuentre en constante estudio.

Este proyecto pretende ser un aporte en torno a la minería de opinión, utilizando los tweets realizado por personas sobre un área específica, obteniéndose esta área a través de los hashtag generados por los usuarios. Esto se llevara a cabo empleando diferentes métodos de clasificación automática de textos como lo son el algoritmo K-NN y K-Means. Los cuáles serán aplicados bajo diferentes contextos en los que se contempla la utilización de diferentes tipos de ponderaciones estadísticas y tres tipos de interpretaciones lingüísticas.

La tesis cuenta con los siguientes capítulos:

En el primer capítulo se presenta el planteamiento de la problemática que aborda la investigación, junto a los objetivos de esta. Además de contener los aspectos metodológicos.

En el segundo capítulo expone los aspectos teóricos relacionados al análisis de sentimientos, la caracterización de los datos y la clasificación automática de textos. A su vez se presenta las ponderaciones a utilizar en el estudio. Además contiene los algoritmos a utilizar en la clasificación, presentando su funcionamiento, ventajas y desventajas.

El tercer capítulo se muestra como serán analizados los resultados, se presentan los conjuntos de datos a utilizar en las experimentaciones y las herramientas empleadas para la implementación de los algoritmos.

En la cuarta sección se entregan los resultados generados por las experimentaciones, con sus respectivas conclusiones.

Como apartado final, se procede a la presentación de un trabajo futuro, para dar pie a la conclusión final de la tesis.

2 Objetivos

2.1. Objetivo General

Este estudio pretende ver la viabilidad de la clasificación automática de documentos a través del Análisis de Sentimiento, utilizando dos algoritmos bajo diversas ponderaciones. Estos algoritmos son K-NN (K Nearest Neighbros) y K-Means y sus contextos de prueba estarán centrados en modificaciones a los diversos conjuntos de datos, estos son, aplicación de TF-IDF y TF-RFL, implementación de un lematizador y la modificación de los documentos mediante N-gramas.

2.2. Objetivos Específicos

- Estudiar y comprender las temáticas con respecto a:
  - Análisis de Sentimiento.
  - Aprendizaje automático.
  - Algoritmo K-NN.
  - Algoritmo K-Means.
  - TF-IDF.
  - TF-RFL.
- Interpretar influencia de las distintas contextos lingüísticos planteados:
  - Lematizador.
  - Bi-grama.
  - Tri-grama.
- Estudiar, implementar y analizar resultados de la aplicación del método de "filtrado de datos" a la ponderación TF-RFL antes de la aplicación de los algoritmos.
- Realizar análisis de resultados obtenidos a través de las diversas metodologías aplicadas con el fin de plantear una interpretación posible a estos resultados.

3 Metodología de Estudio

Esta metodología hace referencia al conjunto de procedimientos o actividades a realizar para alcanzar los objetivos antes descritos.

- Obtener, leer y analizar documentos, estudios o investigaciones referentes a métodos ya aplicados para la clasificación automática de textos.
- Identificar y seleccionar métodos y clasificadores a desarrollar.
- Fase experimental, probar distintos escenarios para comprobar el desempeño de los métodos y clasificadores seleccionados.
- Analizar escenarios propuestos con el fin de identificar los que obtengan mejores resultados.
- Plantear posibles mejoramientos de los procesos con el fin de mejorar los resultados obtenidos.
- Concluir en base a los datos que se obtengan de las aplicaciones de las metodologías, comparando resultado entre ellas.

4 Problemática

La mayoría de las personas gustan de la convivencia en comunidades, en las cuales nos sentimos acompañados, comprendidos y útiles, donde podemos opinar abiertamente de algún tema de interés común, recibir consejos o avisos de importancia. Esto nos permite sentimos más seguros a la hora de tomar las mejores decisiones para nuestras acciones, que pueden ir desde visitar algún lugar o comprar un bien. Antiguamente estas comunidades debían desarrollarse físicamente y el intercambio de opiniones era limitado a la comunidad cercana en la cual se encontraba, pero con los avances de la tecnología, en especial el explosivo incremento del Internet, estas comunidades pasaron a, ser virtuales lo que incrementa este grupo a un nivel global. Esta globalización permite el intercambio de experiencias a un nivel mucho mayor, de forma instantánea y permitiendo un amplia variedad de comunidades e intereses.

Uno de los servicios utilizados para comunicarse por Internet es micro blog, siendo uno de sus propulsores Twitter, el cual permite a la persona enviar texto plano de corta longitud, con un máximo de 140 caracteres, con el tema que el estime. Estos Tweets son publicados en el perfil de usuario y los "seguidores" de la persona pueden visualizarlo y opinar respecto al mismo tema. Las temáticas pueden ir desde conversaciones, noticias, retweets o spam, propagandas, avisos, advertencias y otros más. Esta suerte de Publicación y Seguidor es lo que podemos decir hoy en día una comunidad, en donde existen intercambios de opiniones respecto al tema en cuestión.

Estas opiniones guardan una amplia información respecto a temáticas abordadas por sus creadores, las que podrían contemplar la sensación de bienestar respecto a un tema o el enojo por algún hecho pasado. Pero para obtener esta información se deben analizar estos datos, ya que al ser opiniones generadas globalmente, la obtención manual de la ventaja que entregan las opiniones se hace imposible.

En este proyecto se enfocará en rescatar las diversas opiniones respecto a la temática comercial, es decir, a cualquier tipo de experiencia que pueda ser resumida de forma positiva, neutra o negativa respecto a alguna marca o tienda que el cliente compartió en su perfil. Específicamente se utilizara una herramienta de Twitter llamada "Topsy'sOtterApi" que permite obtener los últimos tweets sobre determinado tema.

Esta información será utilizada por 2 algoritmos de aprendizaje automático, bajo diversas situaciones y contextos, con el fin de analizar posibles resultados.

.

5 Análisis de Sentimiento

Es un campo de la investigación computacional referido al procesamiento de la lengua natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva de algún recurso. También conocido como minería de opinión (opinión mining) o análisis de subjetividad, estos estudios tratan de determinar los ámbitos subjetivos que están implícitos en los contenidos generados por los usuarios, es decir, el tono emocional que hay detrás de una serie de palabras el cual permite entender actitudes, opiniones y emociones expresadas en una oración o texto.

Uno de los usos de estos análisis es en la monitorización de las redes sociales como apoyo a la toma de decisiones de las empresas sobre el rumbo que deben tomar respecto a algún producto. Esto es posible tomando las opiniones generadas por los clientes del producto que a su vez publican en sus redes sociales, en este caso Twitter, de manera que la empresa pueda identificar que decisiones tomar respecto.

Las opiniones se pueden encontrar de forma explícita, argumentando directamente lo bueno o malo, en estos casos se utilizan palabras claves, para calificar determinada situación, como adjetivos; por ejemplo: "La ropa es bonita". En este caso, el adjetivo "bonita" muestra de forma explícita el sentido de la oración. La mayor concentración de estudios se basa en estos tipos de opiniones, debido a lo rápido de identificar la polaridad de la oración.

Otras opiniones son las de carácter implícito, más complejas de identificar, ya que la idea no se expresa de forma directa, sino que, se debe inferir del contexto en el que se encuentre; por ejemplo: "Esperé en la fila como 5 minutos", en esta oración "5 minutos" puede ser tanto bueno como malo dependiendo del contexto en el que se encuentre la opinión.

También, existen opiniones sarcásticas, con un nivel de complejidad aún mayor. Por ejemplo, "Tu eternidad duró un par de meses"; siendo la eternidad un tiempo ilimitado, la promesa finalizó en un tiempo breve en relación a lo esperado.

5.1. Detección de Polaridad

En general encontramos dos tipos de tareas relacionadas con la minería de Opinión, que detallaremos a continuación. [3]

Esta detección se basa en logar determinar o clasificar una opinión como positiva o negativa respecto a una temática. Se debe aclarar que si bien el hecho que una opinión sea positiva no implica que para cualquier ámbito este también lo sea, es decir, en la frase "La ropa de Falabella es barata" podemos apreciar que si analizamos desde un punto de vista de Falabella es una opinión positiva, pero para su competencia es negativa, por ende la

temática y el cómo se abordara la opinión es un factor decisivo en la detección de la polaridad. [3]

Es posible determinar si una opinión es positiva o negativa, pero también puede ser vista dentro de un rango. Existen algunos documentos o frases que se puede apreciar una polaridad mixta, para estos casos se debe hacer una distinción entre la polaridad del sentimiento y la fuerza que este tiene.

Según Mejova [4], para lograr identificar la polaridad de los objetos de estudio primero se debe llevar a cabo la detección del sentimiento, es decir, lograr clasificar el texto como objetivo o subjetivo. Esto se realiza analizando los adjetivos que posee cierta frase, por ejemplo, "El día esta hermoso", en este caso el adjetivo "hermoso" nos entrega de inmediato una clasificación subjetiva explicita ya que podemos identificarla con polaridad positiva fácilmente. [5]

5.2. Análisis del sentimiento basado en características

Este análisis intenta determinar las distintas características o enfoques que se le pueden dar a un determinado producto como por ejemplo la calidad, el precio, la duración de vida, entre otros y a cada una de estas lograr extraer la polaridad de la opinión.

Para realizar el análisis de sentimiento se utilizaran grandes cantidades de datos que contendrán los tweets de los usuarios de la red social. Para hacer la lectura e interpretación de los datos no sería lógico hacerlo de forma manual, por lo que se recurrirá a la Clasificación Automática de Textos.

6 Clasificación Automática de Textos

La clasificación es la acción de ordenar objetos, ideas, palabras según algún criterio o clase dependiendo de las características de lo que se quiera ordenar. Este concepto se puede llevar al ámbito informático y en este caso al análisis de sentimiento para el ordenamiento de opiniones según clases o características que estas tengan.

La clasificación automática de textos también denominada Categorización de Textos o Topic Spotting puede ser definida de dos formas según el límite de las categorías:

- Asignación automática de un conjunto de datos o documentos en una o más categorías preexistentes y ya definidas por el usuario a través de un conjunto de textos pre categorizados sobre los que el sistema lleva a cabo un aprendizaje.
- Asignación de documentos sobre categorías generadas de forma automática por el clasificador. Este tipo no requiere de la intervención del usuario para una preclasificación de sus elementos.

En resumen, la construcción de un clasificador automático de texto comienza con la recopilación y clasificación manual de un conjunto de documentos (documentos de entrenamiento), luego se les aplica algún método de representación adecuada para que finalmente se puedan aplicar los distintos algoritmos de clasificación y así obtener el clasificador.

Para llevar a cabo la clasificación automática de texto se tiene que representar cada documento de los ejemplos de entrenamiento, de manera que a esa representación se le pueda aplicar el algoritmo de clasificación. La representación más utilizada es el modelo vectorial, ésta es manejada ampliamente por los sistemas de recuperación de información.

La gran mayoría de los clasificadores de textos que utilizan algún método de aprendizaje se basan en la inducción probabilística. Otra clase de clasificadores que han experimentado un gran auge en los últimos años, son los simbólicos. Estos se basan en la localización y posterior clasificación de los patrones más representativos del texto y determinantes de cada categoría. [8]

Para implementar un sistema de clasificación se debe:

- Representar los datos
- Implementa Algoritmo de clasificación
- Aplicar Métodos de evaluación

7 Representación de los datos

El primer paso para la clasificación automática consiste en la representación de los conjuntos de datos que se disponen, estos se pueden agrupar en dos tipos principalmente:

- Conjunto de entrenamiento: Utilizados para determinar los parámetros del calificador.
- Conjunto de prueba: Utilizado para estimar el error de generalización.

El objetivo del clasificador es lograr obtener un error de generalización pequeño evitando el sobre-ajuste (sobre-entrenamiento), que consiste en una sobre-valoración de la capacidad predictiva de los modelos, es decir, el algoritmo queda ajustado a características muy específicas de los datos impidiendo así que el sistema no sea capaz de predecir correctamente el resultado en otros casos según lo aprendido con los datos entregados.

El conjunto de entrenamiento se divide a su vez en un sub-conjunto de entrenamiento y otro de validación, el cual es utilizado para ajustar el modelo una vez entrenado el sistema. Se suele utilizar 80% en entrenamiento, 10% en validación y 10% en prueba. Una vez realizada las pruebas, se vuelve a repetir el proceso seleccionando muestras diferentes para obtener resultados más reales.

Figura 7.1 Conjunto de Datos

Para llevar a cabo la extracción de las características de los datos se debe primero tratar el conjunto de datos, a esto se le llama Pre-Procesamiento de los datos.[9]

7.1. Pre- Procesamiento de los Datos

Esta tarea es necesaria para la preparación de los datos que serán utilizados en el análisis. Generalmente es utilizado para la limpieza del ruido con el que vienen los datos.

"El propósito fundamental de la preparación de los datos es la manipulación y transformación de los datos sin refinar para que la información contenida en el conjunto de datos pueda ser descubierta o estar accesible de forma más fácil" [D. Pyle, 1999]

Para poder representar los datos se debe realizar un filtrado de las palabras vacías o también llamadas palabras funcionales ya que no aportan con información relevante, ejemplos de estas son las preposiciones, conjunciones, artículos, pronombres entre otras. Otra estrategia es el uso de un lematizador, el cual tiene como objetivo llevar las palabras a su raíz léxica. Esto se hace con la finalidad de que las palabras que tiene el mismo significado conceptual sean representadas por su raíz conceptual, por ejemplo, caminar, caminará, caminó, caminando se representa con la raíz "camin".

7.2. Caracterización de los datos

La representación más utilizada es el modelo vectorial, ésta es manejada ampliamente por los sistemas de recuperación de información. Este modelo consiste en representar el conjunto de documentos como una matriz de palabras o términos por documentos [Ass K. & Eikvil L., 1999]. Es decir cada documento d_j es representado por medio de un vector:

$$d_j = (W_{1j}, \ldots\ldots, W_{ij})$$

Donde W representa la palabra o termino que puede ir de i=1 a $i = j$ y donde W_{ik} representa un valor numérico que expresa en qué grado el documento d_j posee el termino W_i. Este grado o peso del término se puede calcular de distintas maneras según el enfoque del estudio. [7][8]

7.2.1. N-Gramas

Los textos se pueden representar de variadas formas, la más común es como palabras únicas o Unigrama, es decir nuestro lenguaje normal:

Ayer	Cocine	Fideos	Con	Carne

Figura 7.2 Representación Unigrama

Pero existen otros métodos por los cuales los datos lingüísticos pueden ser procesados, como los son el Bi-grama y el Tri-grama:

Ayer-Cocine	Cocine-Fideos	Fideos-Con	Con-Carne

Figura 7.3 Representación Bi-grama

Ayer-Cocine-Fideos	Cocine-Fideos-Con	Fideos-Con-Carne

Figura 7.4 Representación Tri-grama

7.2.2. Ponderado Booleano

El peso es asignado según se encuentra la palabra en el documento o no.

$$W_{ij} \begin{cases} 1 & \text{Si la palabra aparece} \\ 0 & \text{En caso contrario} \end{cases}$$

7.2.3. Ponderado de frecuencia de término (TF)

El peso es asignado según el número de veces que el término o palabra ocurre en el documento, en otras palabras las veces que ocurre *i* en el documento d_i.

$$W_{ij} = f_{ij}$$

7.2.4. Ponderado por (TF-IDF)

Esta asignación de peso es un producto entre dos medias, la frecuencia de término y la frecuencia invertida de documento. IDF es la proporción inversa entre el número de documentos totales y el número documentos en donde la palabra ocurre al menos una vez.

$$W_{ij} = f_{ij} * \log\left(\frac{N}{n_i}\right)$$

Donde N es el número de documentos totales y n_i es el número de documentos en donde la palabra ocurre al menos una vez. [7]

7.2.5. Ponderado por TF-RFL

La Relevancia de la Frecuencia de una Categoría o Relevance Frequency of a label es una nueva representación para un problema de múltiples categorías propuesto por [10].

$$tf - rfl = tf * \log_2\left(2 + \frac{a_{t,l}}{\max\left(1, \text{mean}(a_{t,\lambda_{j/l}})\right)}\right)$$

Donde la función$\max\left(1, \text{mean}(a_{t,\lambda_{j/l}})\right)$ representa el máximo entre 1 y el promedio de documentos que contienen el término*t* en cualquier categoría diferente de *l*. [10].

El principal objetivo de esta frecuencia es resaltar la importancia que tiene el término a cuantificar sobre las otras clasificaciones.

7.2.6. Ponderado por TF-RRFL

La Relevancia Robusta de la Frecuencia de una Categoría o Robust Relevance Frequency of a label es una segunda preposición de [10] para el problema de múltiples categorías.

$$\mathrm{tf-rrfl} = \mathrm{tf} * \log_2\left(2 + \frac{a_{t,l}}{\max\left(1, \mathrm{median}(a_{t,\lambda_{j/l}})\right)}\right)$$

Donde $\mathrm{median}(a_{t,\lambda_j})$ es la mediana entre las categorías diferentes a *l* que contengan el término *t*.

Para que la clasificación se realice de forma automática se debe implementar un algoritmo de aprendizaje computacional. [10]

7.3. Aprendizaje computacional

La capacidad de aprender se considera como uno de los atributos más distintivos del ser humano y ha sido una de las principales áreas de investigación de la Inteligencia Artificial, cuyo objetico es generalizar comportamientos a partir de información no estructurada suministrada en forma de ejemplos. Se dice que una computadora aprende de una experiencia E con respecto a una clase de tareas T y una medida de desempeño D, si su desempeño en las tareas T, medidas con D, mejora con la experiencia E [Mitchell, 97].

El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis de mercado de valores, reconocimiento del habla, juegos y robótica.

7.4. Taxonomía de sistemas de aprendizaje

Las contribuciones en el aprendizaje computacional o Machine Learning han proliferado y madurado al punto de que es posible hablar ya de una taxonomía de técnicas, las clasificaciones que se usan con mayor frecuencia son aprendizaje supervisado, no supervisado y semi-supervisado.

7.4.1. Aprendizaje Supervisado

Se considera aprendizaje supervisado al algoritmo que aprende de ejemplos ya definidos con anterioridad. Cada ejemplo incluye característica ya definidas en la fase de

Caracterización de los datos, estos suelen ser pares de objetos, normalmente vectores que plasman una etiqueta o clase predefinida por el usuario del algoritmo, de manera tal que cada ejemplo ayuda a diferenciar a que clase pertenece y a cual no.

Con estos datos se construye el modelo o función que es capaz de predecir la clasificación a la que pertenece cualquier objeto de entrada valida que no ha sido visto con anterioridad. La desventaja del modelo es que necesita de una gran cantidad de datos para llevar a cabo el entrenamiento del algoritmo.

Un ejemplo de este modelo son los arboles de decisión, donde a partir de una base de datos se genera un árbol de tal manera que se van haciendo preguntas a medida que avanza por los nodos hijos hasta llegar al último, el cual toma la decisión. Otro ejemplo son los algoritmos por vecindad donde a partir de una muestra para clasificar se decide a que clase pertenece dependiendo de las muestras ya clasificadas con anterioridad que se encuentren más cercanas a ella. [8]

7.4.2. Aprendizaje No supervisado

Se considera aprendizaje no supervisado al algoritmo que no utiliza ejemplos previamente categorizados para su aprendizaje, es decir, el modelo recibe un conjunto de datos aleatorios y los clasifica según patrones.

El conjunto de entrenamiento para estos algoritmos solo incluye atributos, por lo que el modelo debe asociarlos a clases o jerarquías (clustering). [8]

7.4.3. Aprendizaje Semi- Supervisado

Este aprendizaje viene a mejorar una desventaja del Supervisado que sería la gran cantidad de datos necesarios clasificados para que funciones correctamente. Lo que hace este algoritmo es dividir el conjunto de aprendizaje en dos, uno ya clasificado que viene siento lo mismo que en el caso del supervisado y otro grupo con datos no clasificados.

El objetivo es utilizar los datos clasificados para aprender a clasificar los datos que no se encuentran etiquetados. Algunos ejemplos de estos son Co-training, Assemble y self-training. [5]

7.5. Tipos de aprendizaje

Existe una división diferente a la anterior en la que se basa en las formas de aprendizaje:

- **Aprendizaje Analítico o Deductivo**: se aplica la deducción para obtener descripciones generales a partir de un ejemplo. Se basa en el razonamiento deductivo, obtener conocimiento mediante el uso de mecanismos bien establecidos. Este conocimiento no es nuevo (está implícito). Nuevo conocimiento no invalida el ya obtenido.
- **Aprendizaje Analógico**: se buscan soluciones a problemas nuevos basándose tratando de encontrar similitudes con problemas ya conocidos, adaptando la solución.
- **Aprendizaje Genético**: aplica algoritmos inspirados en la teoría de la evolución para encontrar descripciones generales a conjuntos de ejemplos.
- **Aprendizaje Conexionista**: busca descripciones generales mediante el uso de la capacidad de adaptación de redes de neuronas artificiales.
- **Aprendizaje Inductivo**: el aprendizaje inductivo es la capacidad de obtener nuevos conceptos, más generales, a partir de ejemplos. Este tipo de aprendizaje conlleva un proceso de generalización/especialización sobre el conjunto de ejemplos de entrada. Los algoritmos implementados son, además, incrementales, es decir, el procesamiento de los ejemplos se realiza uno a uno. Esta característica, permite visualizar el efecto causado por cada uno de los ejemplos de entrada, en el proceso de obtención del concepto final. Además de la generalización de conceptos, el programa permite clasificar conjuntos de ejemplos a partir de los conceptos obtenidos anteriormente. De este modo, se puede comprobar, para cada ejemplo de un conjunto dado, a qué clase pertenece dicho ejemplo. Este aprendizaje se puede dividir en los anteriores mencionados: los algoritmos supervisados y los no supervisados. [8]

8 Aprendizaje Inductivo Supervisado

8.1. KNN o K-vecinos más cercanos

El método KNN (K Nearest Neighbros) es un algoritmo de clasificación supervisada basada en un conjunto de entrenamiento. Las reglas de clasificación por vecindad están basadas en la búsqueda de un conjunto de prototipos ya clasificados que se encuentran más cercanos al elemento a clasificar.

El coste del aprendizaje es 0, todo el coste pasa al cálculo de la predicción. Se conoce como mecanismo de aprendizaje perezoso (lazy learning).

Para llevar a cabo la clasificación primero se debe especificar una métrica para poder medir la proximidad entre los vecinos y el elemento a clasificar, generalmente es utilizado la distancia euclidiana.

Este método se basa en la suposición de que la clase del patrón a etiquetar, *X*, es la del conjunto más cercano a X_{NN} (conjunto de referencia). Si $K_i(X)$ es el número de muestras de las clases presentes en los *k* vecinos más próximos a *X*, entonces:

$$d(X) = w_c \text{ si } K_c(X) = \max_{i=1\rightarrow j} K_i(K)$$

8.2. Elección de la variable K

La elección de K es un tema delicado, de esta depende mucho los resultados y la calidad de estos. La mejor elección de K depende fundamentalmente de los datos; generalmente valores grandes de K reducen el efecto de ruido en la clasificación, pero crean límites entre las clases parecidas.

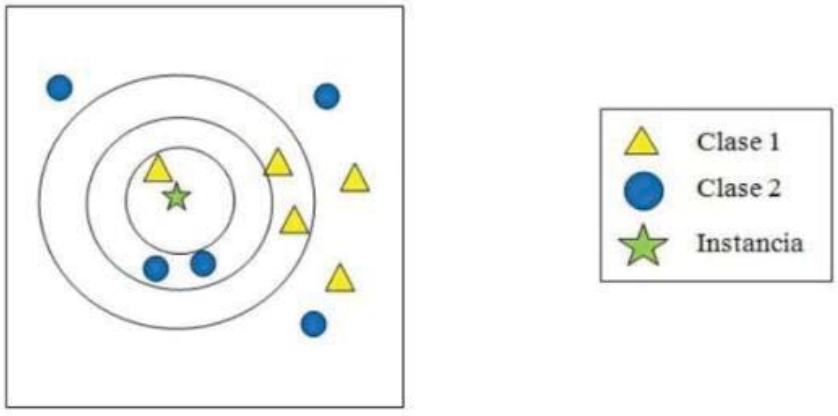

Figura 8.1 Representación de K

Supongamos un espacio de representación bidimensional y una serie de prototipos de diversas clases especificados en él. Dado un patrón cualquiera X, se debe considerar los k prototipos más próximos a X, estos estarán localizados en un círculo centrado en X. En la figura 8.1 se muestra 3 k diferente. Se puede apreciar que en la primera instancia de un K=1, la instancia sería clasificada como clase 1. Para la segunda instancia k=3, se tiene 2 clases distintas, pero predomina la clase 2, por lo que la instancia será categorizada como tal. Para la instancia donde k=3, siguiendo la misma lógica anterior, será clasificada como clase1.

Esto demuestra la importancia que cumple la determinación de K y cómo influye esta variable en el algoritmo de clasificación. Se debe tener en consideración que el área del circulo va aumentando según la variable K y que en regiones más densamente pobladas, este crecimiento es menor que en regiones donde los puntos están más dispersos.

Se debe tener en consideración:

- Si K es muy pequeño, el modelo será muy sensible a puntos o variables que son atípicos o que son ruido.
- Si K es muy grande, el modelo tiende a asignar a la clase más grande.(figura 8.2)

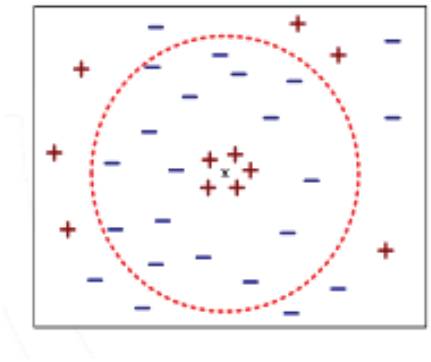

Figura 8.2 K grande

8.3. Algoritmo KNN

El proceso de aprendizaje de este clasificador consiste en almacenar en un vector el conjunto de entrenamiento, junto a la clase asociada a cada muestra de este conjunto (datos ya clasificados). Luego se debe calcular la distancia entre cada muestra de entrenamiento y el vector a clasificar, seleccionando las K muestras más cercanas. Luego se debe hacer el mismo proceso pero con los datos de validación, para así diseñar el clasificador. Luego se calcula el porcentaje de clasificación para poder conocer el poder de generalización. [9]

```
Comienzo
     Entrada: D {(X1, C1),..., (XN, CN)}
            X= (x1,..., xn) nuevo caso a clasificar
     Para todo objeto ya clasificado (xi,ci)
            Calcular di=d (xi, x)
     Ordenar di (i=1,…, N) en orden ascendente
     Seleccionar los K casos D_X^K ya clasificados más cercanos a x
     Asignar a x la clase más frecuente en D_X^K
Fin
```

8.4. Ventajas y Desventajas del Algoritmo K-NN

Ventajas:

- El coste del aprendizaje es nulo.
- No se necesita hacer suposición alguna sobre los conceptos a aprender.
- Se puede aprender conceptos complejos usando funciones sencillas como aproximaciones locales.
- Es muy tolerante al ruido.

Desventajas:

- El coste de encontrar los k mejores vecinos es grande.
- No hay un mecanismo para decidir el valor óptimo para k(depende de cada conjunto de datos).
- Su rendimiento baja si el número de descriptores crece.

9 Clustering

Los algoritmos de segmentación (también conocidos como algoritmos de agrupamiento o, en inglés, clustering) pertenecen al grupo de métodos de minería de datos (data mining) definido como no supervisados. El objetivo del clustering no es clasificar, estimar o predecir una variable, sino entender la estructura macroscópica y relaciones entre objetos, considerando las maneras en las que estos son similares y diferentes. En otras palabras, se enfoca en segmentar el conjunto completo de datos en subgrupos homogéneos.

Clustering posee varias características, de las cuales destacan:

- Escalabilidad: estos algoritmos funcionan en situaciones donde se tengan pocos datos, como en otras con muchos datos.
- Clusters de formar arbitraria: según la función que se le aplique para la diferenciación, es la forma que puede tomar el cluster. En los que están basados en distancias numéricas tienden a encontrar clusters esféricos.
- Capacidad de manejar diferentes tipos de atributos: numéricos (lo más común), binarios, nominales, ordinales, etc.
- Capacidad de añadir restricciones: permiten la incorporación de diversas restricciones con el fin de obtener mejores resultados, ya sea limitando la agrupación, seleccionando diversas formas de medición o incorporando nuevas variables.
- Manejo de ruido: muchos de estos algoritmos son sensibles a datos erróneos.
- Pueden funcionar eficientemente con datos de alta dimensionalidad.
- Son independiente del orden de los datos.
- Los clusters son interpretables y utilizables.

Clustering trata, fundamentalmente, de resolver el siguiente problema: dado un conjunto de individuos (de N elementos) caracterizados por la información de n variables X_j, (j = 1,2,..., n), se plantea que el algoritmo sea capaz de clasificarlos de manera que los individuos pertenecientes a un grupo (cluster) sean tan similares entre sí como sea posible, siendo los distintos grupos entre ellos tan disimilares.

Con el análisis cluster se pretende encontrar un conjunto de grupos a los que ir asignando los distintos individuos por algún criterio de homogeneidad. Por lo tanto, se hace imprescindible definir una medida de similitud o bien de divergencia para ir clasificando a los individuos en unos u otros grupos.

Existes diversos métodos de clustering:

- Jerárquicos: los datos se agrupan de manera arborescente (top-dow o bottom-up).
- No Jerárquicos: generar particiones a un solo nivel (K-means).
- Paramétricos: se asume que las densidades condicionales de los grupos tienen cierta forma paramétrica conocida y se reduce a estimar los parámetros. (Algoritmo EM).

- No Paramétricos: no se asume nada sobre el modo en el que se agrupan los objetos. [14]

9.1. K-Means

El algoritmo K-means, creado por MacQueen en 1967 es el algoritmo de clustering más conocido y utilizado ya que es eficaz a pesar de su simple aplicación. Sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado número K de clústeres, previamente determinado. El nombre de K-means viene porque representa cada uno de los clusters por la media de sus puntos, es decir, por su centroide. La representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. Cada cluster por tanto es caracterizado por su centro o centroide que se encuentra en el centro o el medio de los elementos que componen el cluster. [9]

9.2. Algoritmo K-Means

- **Etapa 1**: elegir aleatoriamente K objetos que forman así los K clusters iniciales. Para cada cluster k, el valor inicial del centro es xi, siendo este, el único objeto perteneciente al clusters.
- **Etapa 2**: se debe reasignar los objetos del cluster. Para cada objeto x, el prototipo que se le asigna es el que es más próximo al objeto, según una medida de distancia, (habitualmente la medida euclidiana).
- **Etapa 3**: una vez que todos los objetos son colocados, recalcular los centros de K cluster. Estos nuevos cluster representan la media del total de objetos asignados al cluster.
- **Etapa 4**: repetir las etapas 2 y 3 hasta que no se hagan más reasignaciones. Aunque el algoritmo termina siempre, no se garantiza el obtener la solución óptima. En efecto, el algoritmo es muy sensible a la elección aleatoria de los K centros iniciales. Esta es la razón por la que, se utiliza el algoritmo del K-Means numerosas veces sobre un mismo conjunto de datos para intentar minimizar este efecto, sabiendo que a centros iniciales lo más espaciados posibles dan mejores resultados. [9]

9.3. Elección de los K objetos y sus centroides

La determinación de K es la siguiente:

- Si K es muy pequeño, se agruparan grupos “distintos”.
- Si se elige un K muy grande, hay centros que pueden quedar huérfanos, o sin agrupación.
- El valor de K puede determinarse según alguna heurística.

Por consiguiente, para lograr un K óptimo o una aproximación concluyente, se optará por realizar varias pruebas con los datos, para así al analizar los resultados, lograr estimar de mejor manera la variable K.

10 Distancias estadísticas

Hasta el momento en los algoritmos antes mencionados se ha supuesto la utilización como medio de cálculo de distancia solo el método euclidiano, pero a continuación se mostraran los diversos métodos a aplicar.

Primero se debe especificar que para que un cálculo de distancia sea considerado como método, está debe cumplir ciertas propiedades.

Sea δ una distancia sobre los puntos (i, j): [15]

- P. 1 : $\delta_{ij} \geq 0$
- P. 2 : $\delta_{ii} = 0$
- P. 3 : $\delta_{ij} = \delta_{ji}$
- P. 4 : $\delta_{ij} = \delta_{ik} + \delta_{jk}$

10.1. Distancia Euclidiana

El método euclidiano es el más utilizado y fácil de comprender. Sea R^n el espacio euclidiano de n dimensiones, cuyos elementos son las n-uplas ordenadas de números reales. Dados dos objetos I_1 y I_2 medidos según n variables (X_1,…, Xn):

$$d_{I_1 I_2} = \sqrt{\sum_{k=1}^{n} (X_{1k} - X_{2k})^2}$$

10.2. Distancia Euclidiana Normalizada

La distancia euclidiana, a pesar de su sencillez de cálculo y de que verifica algunas propiedades interesantes tiene dos graves inconvenientes:

- Es sensible a las unidades de medida de las variables: las diferencias entre los valores de variables medidas con valores altos contribuirán en mucha mayor medida que las diferencias entre los valores de las variables con valores bajos. Como consecuencia de ello, los cambios de escala determinarán, también, cambios en la distancia entre los individuos.
- Si las variables utilizadas están correlacionadas, nos darán una información, en gran medida redundante.

Para dar solución a estos inconvenientes esta la distancia euclidiana normalizada:

$$d(i,j) = (X_i - X_j)' S^{-1} (X_i - X_j)$$

Donde S es un matriz diagonal con las varianzas en su diagonal principal y ceros en el resto de sus elementos.

Empleando este tipo de distancia solventamos el inconveniente de los efectos de unidades de medida distintas de las variables y obtenemos una distancia que no dependerá de las unidades de medida.

Sin embargo, la alta correlación entre algunas variables puede seguir siendo un grave inconveniente. [18]

11 Análisis de resultados

Para poder realizar un correcto análisis de los resultados se deben establecer medidas con las cuales se podrán realizar comparativas entre los distintos algoritmos y sus respectivas iteraciones.

Conceptos a definir para llevar a cabo las medidas:

- Grado de exactitud: Es el grado de concordancia entre las clases asignadas por el clasificador y sus ubicaciones correctas según los datos recolectados por el usuario y considerados como datos de referencia a tomar. [11]
- Matriz de Confusión: Es la herramienta más utilizada para la estimación de exactitud de un clasificador, también llamada matriz de error o de contingencia. Esta es una matriz cuadrada de n x n, donde n es el número de clases. Dicha matriz muestra la relación entre las series de medidas correspondientes al área en estudio. En una matriz de confusión las columnas corresponden a los datos de referencia, mientras que las filas corresponden a las asignaciones del clasificador. [11]

Para la realización de los análisis se puede considerar una división de estas según aplicaciones realizadas a los algoritmos planteados con anterioridad y las contextualización de los datos a utilizar por estos:

- K-NN: Para este algoritmo se plantearon dos grandes grupos de análisis:
  - Según K, es decir, según la cantidad de vecinos cercanos, se sugirió utilizar 3 diferentes: 1, 2,5 vecindades. El principal objetivo es identificar como influye la vecindad en los respectivos análisis deseados.
  - Según porcentaje de datos de entrenamiento: Al ser un algoritmo que necesita de un grupo previo de entrenamiento, se plantea la experimentación según la cantidad de datos utilizados para entrenar al algoritmo versus los que debe clasificar. Estos porcentajes fueron de 95%, 90%,80%, 70% y 60% 50%. A través de este experimento se puede analizar como la cantidad de datos de entrenamiento afecta al resultado, considerar que mientras más alta sea el porcentaje de entrenamiento, más pequeño es el conjunto de datos a clasificar.
- K-Means: Para este algoritmo se planteó la comparativa entre los porcentajes de acertados entre cada contexto en el que se puso a prueba el algoritmo.
- Contextualización de datos: Se plantearon 3 factores de contextos distintos con el fin de comprender como afecta los datos de entrada al algoritmo en los resultados finales de estos, permitiendo comparar cual metodología podría predominar sobre otra estas fueron:
  - Lematizador
  - Bi-grama
  - Tri-grama

Con el fin de mejorar el rendimiento de la ponderación TF-RFL, se ha incorporado un método, la filtración de datos. Este método consiste en la eliminación de un sector de la gama de datos, con el fin de reducir atributos que lleven a redundancia a la hora de aplicar el algoritmo clasificador. Esta actividad se realiza después de la ponderación, cuando los atributos ya poseen sus respectivos pesos según la categoría a la que pertenecen. El filtrado es aplicado según alguna razón escogida y elimina los datos mínimos que comprenden dentro de esa razón, es decir, si se posee una gama de 1.000 atributos y se escoge una razón de filtrado del 10%, entonces la eliminación comprenderá los primeros 100 datos más pequeños, por lo que el algoritmo utilizaría el 90% de los datos con mayor peso.

Se optó por 3 métodos:

- Desviación estándar
- Filtrado del 10% datos
- Filtrado del 5% de datos

Cabe señalar que aumentar el porcentaje de filtrado representa una eliminación masiva de datos, por lo que se debe considerar que el 10% comprender una gran cantidad de datos.

12 Corpus

Para una mejor interpretación de los datos se utilizó dos tipos de corpus o "Conjunto de datos" con el objetivo de diversificar las respuesta entregadas por parte de los algoritmos, dando pie a una mejor y más formal entrega de resultados. Ambos corpus son de tipo lingüístico lo que significa que están compuestos por un número de conjuntos de palabras con una finalidad cada una con una finalidad concreta, ya sea informar o entregar opinión.

Uno de los corpus corresponde a 3.000 tweets de opinión generados por usuarios a través de la plataforma Twitter. Estas opiniones están enfocadas en una de las compañías más grandes de retail en Chile, "Falabella". Estas opiniones se encuentran divididas equitativamente según polaridad, es decir, 1.000 opiniones positivos, 1.000 opiniones negativas y 1.000 neutras. Estos datos fueron obtenidos a través de la herramienta "Topsy's OtterApi", la cual permite descargar tweets referentes a un determinado tema o hashtag.

El segundo corpus utilizado en el estudio es "Reuters-21578", este conjunto de datos contiene una colección es de 21578 artículos publicados por una importante agencia de noticias perteneciente al Reino Unido. Para la experimentación se utilizaron los datos de la categoría: Fusiones y Adquisiciones (Mergers/Acquisitions, ACQ). Estos datos vienen categorizados de forma binaria (positiva y negativa), es decir, cada conjunto provee datos tanto para cuando la noticia pertenece o no a esa categoría, específicamente ACQ contiene 10.528 instancias negativas y 2369 positivas.

12.1. Contexto de Experimentación

La experimentación se realizó en dos fases, primero se utilizaron los datos de los tweets, para los cuales se les aplico una serie de medidas que se profundizaran más adelante, obteniendo así diversos resultados, los cuales fueron utilizados como guía para la realización de la segunda fase, la experimentación con el corpus Reuters-2158. En ambos casos se utilizó la herramienta WEKA.

12.1.1. WEKA

El software Weka (Waikato Enviromnet for Knowledge Analysis) es un conjunto de algoritmos de aprendizaje automático, el cual puede ser utilizado para tareas de minería de datos. Los algoritmos o bien se pueden aplicar directamente a un conjunto de datos o llamadas de su propio código Java. Weka contiene herramientas para el procesamiento previo de datos, clasificación, regresión, clustering, reglas de asociación, y la visualización. También es muy adecuado para el desarrollo de nuevos sistemas de aprendizaje de máquina.

Además Weka es un software libre y de código abierto, por lo que es muy versátil para la continuación o aplicación de nuevos estudios. La versión utilizada fue Weka 3.6.9 [28]

12.2. Pasos de la experimentación

Como se explicó anterior mente, la experimentación consto de dos fases:

12.2.1. Experimentación con primer Corpus (Tweets):

- Recepción y adaptación de los datos: Se reciben los datos y se modifican de manera tal que Weka pueda interpretarlos.
- Aplicación y obtención de formatos lingüísticos: formatos en Bi-gramas y Trigramas a través Java, como medio de reducción de ruido.
- Aplicación de lematizador: Se aplica un lematizador ajeno al que cuenta Weka, el cual se encuentra en código Java. La idea principal es averiguar cómo influye la reducción del ruido en la generación de resultados.
- Aplicación de ponderación: Se aplica la ponderación TF-RFL a los datos, esto se realiza a través de un software creado en Java, el cual calcula los pesos de los datos y los entrega para ser interpretados por Weka.
- Generación de datos: Se importan los datos a Weka y se selecciona los algoritmos a utilizar, en este caso serán K-NN, con vecindades en 1,2 y 5, y K-Means con un número de clustering de 3. Además de generar reiteradas instanciaciones según la cantidad de datos de entrenamiento o filtros a realizar.
- Interpretación y análisis de datos.
- Conclusión de experimentación: La fase más importante, en la cual se reúnen todos los datos y se generan conclusiones.

12.2.2. Experimentación con segundo Corpus (Reuters):

- Obtención de los conjuntos de datos vía web, estos se encuentran libres para su uso.
- Adaptación del conjunto de datos para posterior aplicación de ponderación TF-RFL.
- Generación de datos: Utilización del software Weka bajo el algoritmo K-NN con vecindades en 1, 5,10 y 30. Se incorporó además un corpus de datos filtrados mediante el 5% de los datos más bajos luego de aplicado la ponderación TR-RFL.
- Interpretación y análisis de datos generados.
- Generación de conclusión respecto a los datos obtenidos.

13 Resultados obtenidos (Twitter)

Los resultados fueron divididos según los algoritmos planteados con anterioridad, a cada uno de estos se les uso en contextos de Lematización y aplicaciones de N-Gramas (Bi-grama y Tri-grama).Los valores numéricos en tablas pueden ser encontrados en los anexos.

13.1. Algoritmo K-NN

El primer caso de estudio comprende al algoritmo K-NN, para el cual se dividieron los resultados según Lematizador para mejor comprensión.

13.1.1. K-NN Sin Lematizar

En la figura 13.1 se puede apreciar el promedio según vecindad, es decir, el promedio de resultados entregados bajo la vecindad K=1, K=2 y K=5 versus los diferentes grupos de entrenamientos asignados. A su vez se logra visualizar el rendimiento según el conjunto de entrenamiento considerado para el algoritmo.

El desglose de cada vecindad respecto al grupo de entrenamiento asignado se encuentra en el anexo [A.1].

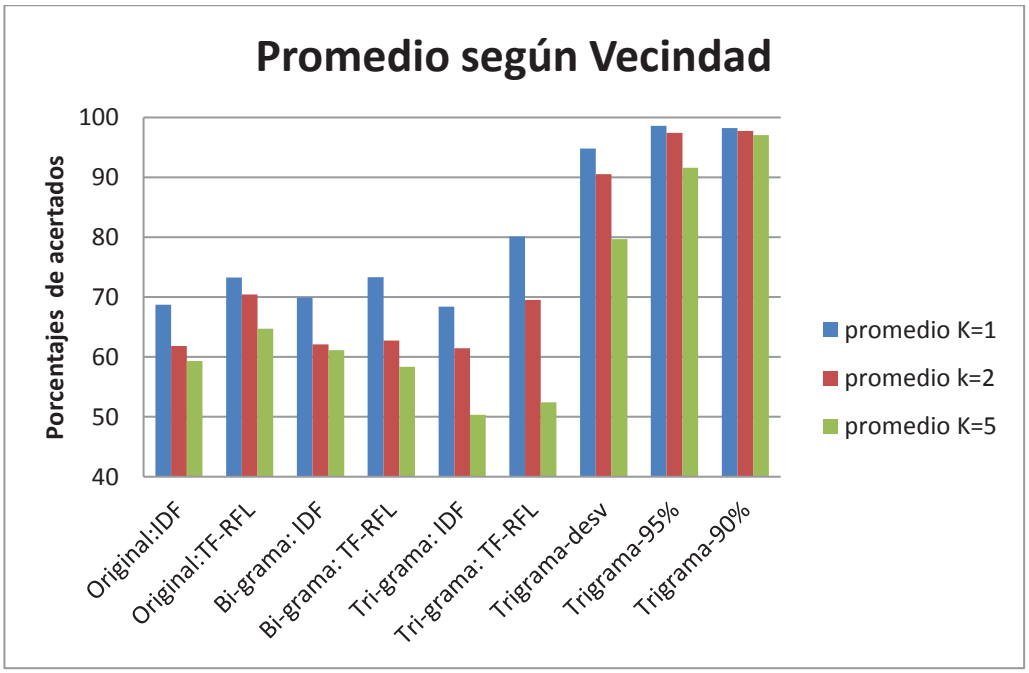

Figura 13.1 Promedio K-NN sin Lematizar (Vecindad)

Como se logra apreciar, el algoritmo plantea una muy clara diferenciación de rendimiento según el tratado de los datos, considerando el Tri-grama: IDF como el peor rendimiento bajo la vecindad K=5. A su vez, cabe señalar que el tratado de datos según filtración de resultados bajo la ponderación TF-RFL aumentó considerablemente los porcentajes de acertados, lo que queda demostrado en la curva observada. Además se aprecia que el número de K influye de una manera determinante, específicamente se puede visualizar que a medida que K aumenta disminuye su porcentaje de asertividad, por lo que K=1 es el más óptimo observable.

13.1.2. K-NN Lematizado

El grafico 13.2 comprende los porcentajes generados por K-NN en un contexto de datos Lematizados. Como se visualiza, la lematización logro una estabilidad de los datos, permitiendo una visualización clara respecto a los porcentajes de acertados, a diferencia del grafico 13.1. Sin embargo, destacan tres alzas considerables en los porcentajes, las cuales comprenden justamente al ponderado TF-RFL con un máximo en los filtros realizados a esta ponderación. Por el contrario, las bajas más destacadas resultan ser en la ponderación IDF.

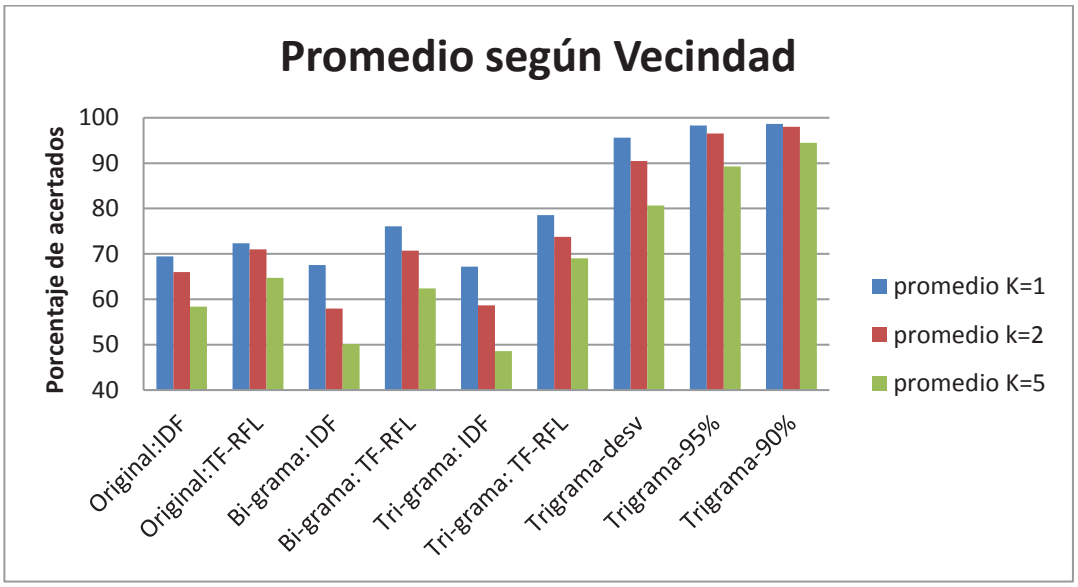

Figura 13.2 Promedio según vecindad K-NN (Lematizado)

13.1.3. Comparación de resultados según Lematización (K-NN)

Como anteriormente se consideraron los resultados por separado entre Lematizado y sin Lematizar, en el grafico 13.3 se puede apreciar una comparativa entre los promedios obtenidos en las figuras 13.1 y 13.2, en el cual se aprecia con más claridad la diferencia entre las alzas bien marcadas en el promedio Lematizado, pero además, se puede observar que no existe diferencia considerable entre los porcentajes de acertados obtenidos.

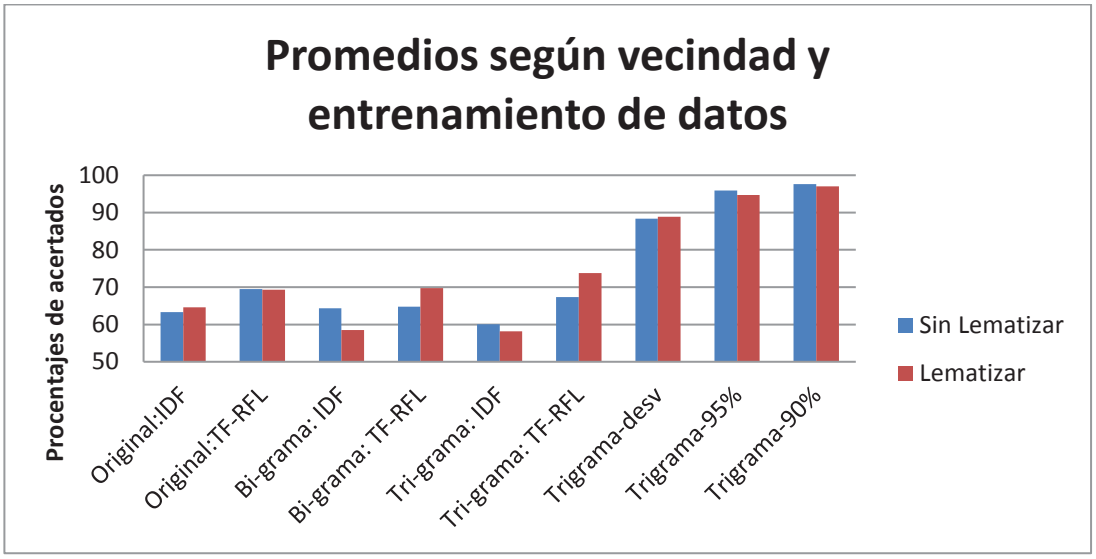

Figura 13.3 K-NN promedio Lematizado vs Sin Lematizar

13.2. Algoritmo K-Means

El segundo caso de estudio comprende al algoritmo K-Means, para el cual no hubo necesidad de realizar una división de su análisis como en el caso anterior, los datos para la generación de este grafico se encuentra en el anexo [B].

En la figura 13.4 se puede apreciar que el rendimiento del algoritmo está muy por debajo de su par el K-NN, pero al igual que este, se puede apreciar un aumento de resultados para los factores en los cuales se realizó el filtrado de datos, llegando a un máximo en el filtrado del 90% de un 53.81% de asignación correcta.

A su vez, se puede apreciar que no existe mayor diferenciación entre si los datos fueron o no Lematizados, por lo que da a entender que esta propiedad de los datos no afecta en gran medida a este algoritmo.

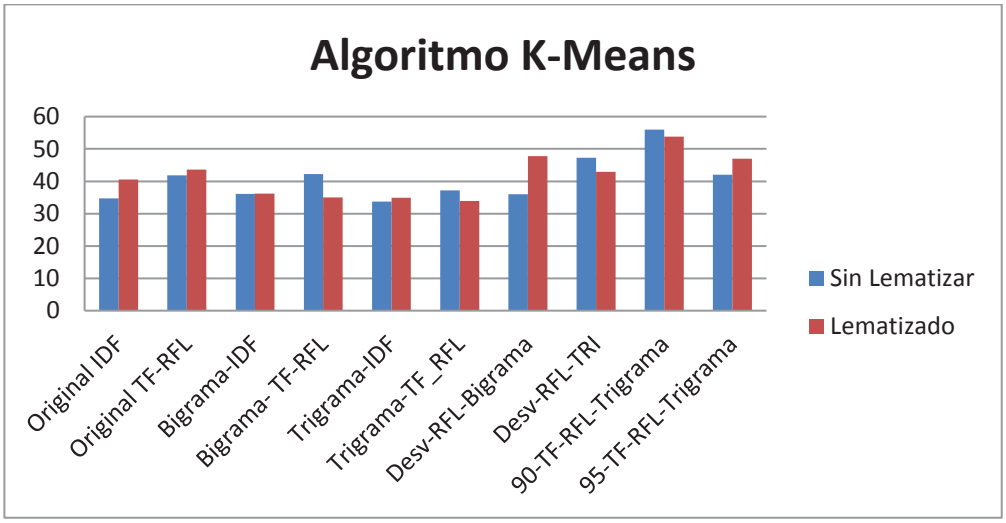

Figura 13.4 K-Means comparativo

13.3. Conclusión primera experimentación

Los resultados entregados por Weka a través de los algoritmos K-NN y K-Means permiten generar varias conclusiones a partir de los modos lingüísticos utilizados, las diferencia entre los resultados según la variable K y las ponderaciones IDF y TF-RFL.

En los gráficos 13.1 y 13.2 se puede apreciar que la aplicación de una ordenanza lingüística diferente no presenta una mejora significativa en los resultados, es decir, los resultados generados por los N-Gramas no presentan un aumento en el porcentaje de asertividad.

Otro punto a apreciar es la significativa diferencia entre las ponderaciones IDF y TF-RFL. En los gráficos anteriores se puede observar el aumento significativo en el promedio porcentual equivalente a un 11%, por lo que se puede concluir que la ponderación TF-RFL logra mejores resultados que IDF.

Un factor muy importante a destacar es la variedad de resultados generados según los diferentes valores que tomó K. Tanto en 13.1 como en 13.2 se aprecia una significativa mejora de los datos a medida que disminuye K llegando al más óptimo con K=1.

Respecto al algoritmo K-Means, este no entrega un porcentaje de asertividad concluyente para poder generar un estudio más acabado. Su porcentaje máximo fue del 53,8%, por lo que la utilización de este método bajo las características antes citadas no es recomendable por este estudio.

14 Resultados obtenidos (Reuters)

Esta experimentación fue realizada en base a los resultados obtenidos anteriormente, por lo que no se aplicó el algoritmo K-Means ni las interpretaciones lingüísticas. La obtención de estos datos se centró en la aplicación de las metodologías que más aumentaron el porcentaje de asertividad del método K-NN. Para la obtención de resultados se realizó por cada contexto 3 iteraciones mezclando el conjunto de datos antes de la distinción entre conjunto de entrenamiento y prueba, esto con la finalidad de conseguir resultados más fidedignos.

El desglose respectivo a esta experimentación se encuentra detallado en el anexo [A.E], en el cual se puede apreciar tablas según los promedios generados en los tres conjuntos de datos empleados para cada contexto a experimentar.

14.1. Algoritmo K-NN

Para esta segunda fase de experimentación se emplearon los datos del conjunto de Fusiones y Adquisiciones (Mergers/Acquisitions, ACQ), para lo cual se aplicaron 3 diferentes ejes de experimentación, partiendo por diferentes tipos de ponderaciones (IDF, TF-RFL, TF-RFL al 5% y al 10%), variación en la constante K (1, 5, 10,30) y variación en el porcentaje de entrenamiento y pruebas (50%,60% y 70%).

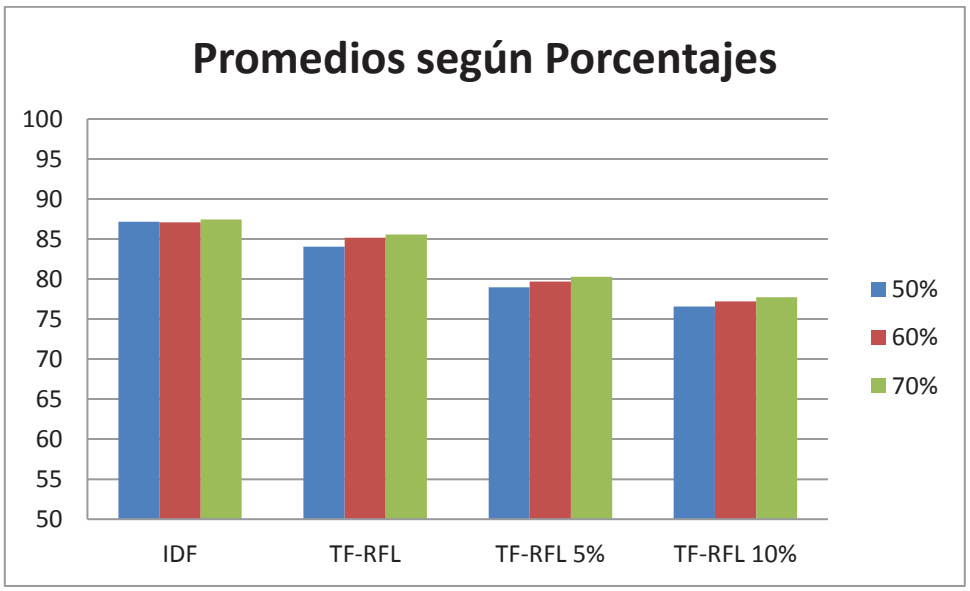

Figura 14.1 Promedios Ponderación y porcentajes

Para comenzar se presentan los datos generados por parte del algoritmos K-NN para las diferentes ponderaciones aplicadas en contraposición de los porcentajes de entrenamientos escogidos. Se puede apreciar una disminución en torno al 3% entre las características IDF y TF-RFL disminuyendo a medida que se aplica o aumenta el filtro de datos. Otro factor importante a destacar es la leve diferencia de resultados por parte de los porcentajes aplicados a los conjuntos de datos de entrenamiento. Se aprecia una diferencia promedio del 0,7% lo cual revela el poco impacto que tiene esta característica en este conjunto de datos.

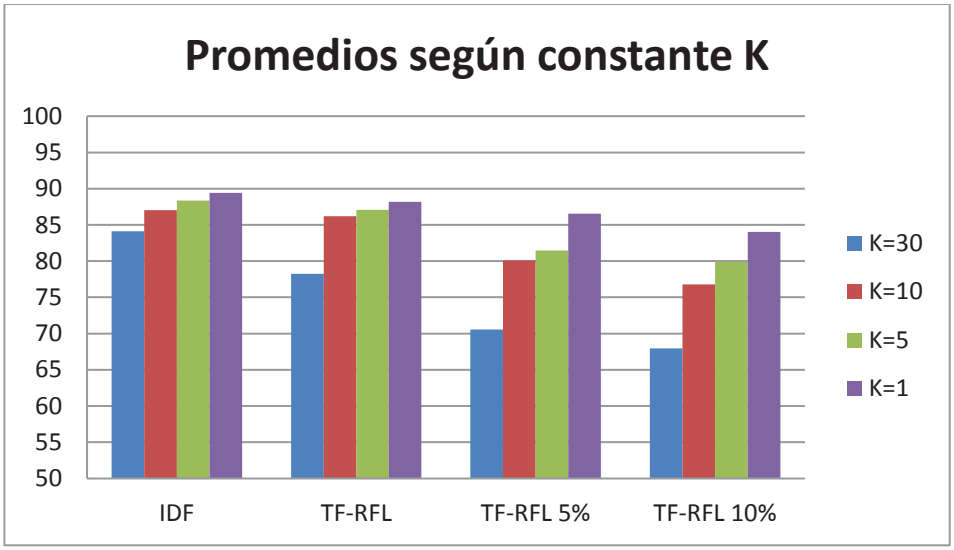

Figura 14.2 Promedios Ponderación y constante K

En la figura 14.2 se puede apreciar nuevamente las diferentes ponderaciones aplicadas según los tipos de valores que se le ha otorgado a K. Partiendo de lo observado en la figura 14.1 en cuanto a la disminución promedio de los resultados por parte de las ponderaciones, se suma a estos la notoria diferencia entre los datos entregados según la constante K. A medida que el valor de K disminuye aumenta considerablemente los resultados obtenidos, llegando a un máximo promedio 11,8%. El valor máximo generado fue de 89,7% cuando la variable K toma valor 1 en la ponderación IDF y con el 70% del conjunto total usado para entrenamiento.

Para poder apreciar con mayor claridad esta diferencia véase la figura 14.3, en la cual se muestra el promedio generado por K-NN para los distintos valores que toma K.

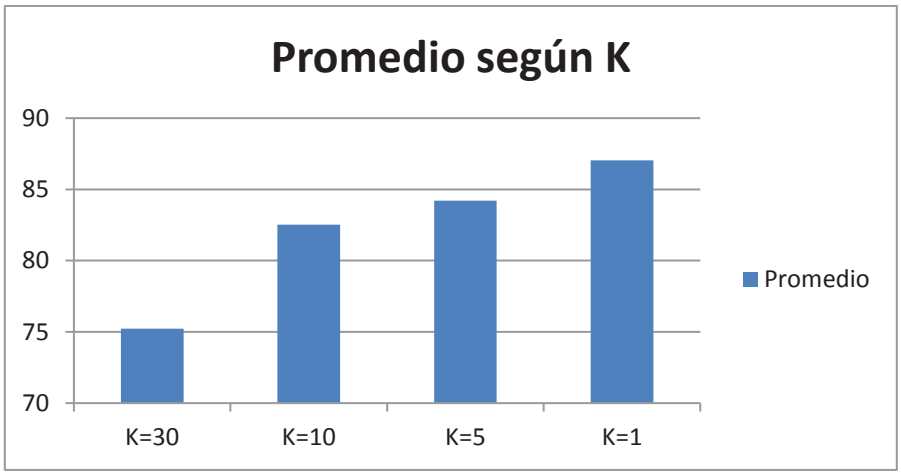

Figura 14.3 Promedio según K

14.2. Conclusión segunda experimentación

A partir de los resultados entregados por WEKA para esta segunda experimentación en la que se analizó el conjunto de datos "Reuters-21578", se puede desprender varios puntos a destacar:

- Mínima mejora en los resultados al modificar el porcentaje de datos para entrenamiento lo que supondría un buen indicio ya que permite obtener buenos resultados no necesariamente con grandes cantidades de datos para entrenar, lo que habla de la eficiencia del algoritmo.
- Mejora de resultados a medida que disminuye la variable K, lo que implica que las ponderaciones aplicadas, ya sean IDF o TF-RFL logran potenciar la diferenciación entre cada categoría (positiva y negativa) una vez aplicadas, lo que ayuda a mejorar el rendimiento del algoritmo K-NN.
- Destacar el descenso del rendimiento al cambiar de ponderación, esto debido a que TF-RFL está pensado para entregar mejores resultados cuando se procesan conjuntos de datos de más de 2 categorías, como es el caso de la experimentación con los Tweets (Positivo, Negativo y Neutro). Esto queda demostrado en la ecuación de la ponderación TF-RFL entregada en el puto 7.2.5, en la cual podemos observar que en el divisor de la ecuación, en el procesos de la elección de un máximo, este debe elegir solo entre 2 categorías, por lo que en el caso de no existir esta palabra en la categoría contraria nos da como única opción el valor 1.
- Se puede observar una baja en el porcentaje de asertividad en cuanto a la aplicación del filtrado de datos para la ponderación TF-RFL, esto se debe a la constitución de los datos utilizados en la experimentación. El hecho de que los datos consten de solo dos categorías, sumado a la gran diferencia de instancias entre estas (2369 positivos, 10528 negativos), es que se genera esta baja en los resultados. Al aplicar un filtrado en los datos se genera una eliminación de estos

según la razón escogida por lo que al usar este método en la categoría positiva, esta al ser minoritaria, pondrá en dificultad a la mayoritaria (negativa) al disminuir aún más los datos a comprar para la realización de la clasificación.

15 Trabajo Futuro

Se pueden considerar diversos trabajos Futuros respecto a este estudio, considerando factores mejoras en los algoritmos o consolidación de métodos empleados en ambas experimentaciones presentadas con anterioridad, dentro de los trabajos futuros a considerar se encuetan:

- Mejora en el algoritmo K-Mean ya sea en el estudio de las semillas a utilizar o en la incorporación de nuevos métodos que den origen a mejores resultados.
- Para el algoritmo K-NN y según las conclusiones entregadas con anterioridad en ambas experimentaciones, se podría dar paso a una nueva puesta a prueba, esta vez considerando un estudio más amplio en el que se consideren datos de opinión y con mayor diversidad de categorías. Esto con el fin de verificar los datos obtenidos en este estudio.
- A su vez, la utilización del método de "filtración de datos" utilizado para la ponderación TF-RFL en el algoritmo K-NN, dio resultados positivos para la clasificación de datos nivelados en cuanto a instancias y que poseen más de dos categorías como fue el caso de la experimentación con Twitter. Por lo que una próxima experimentación con datos de opinión que contemplen más de 2 categorías podría llevar a consolidar este método como uno de los factores a considerar en la aplicación de la ponderación TF-RFL

16 Conclusión

Actualmente, el avance tecnológico ha permitido a los usuarios del internet pertenecer a una red global a la cual pueden acceder en cualquier momento y lugar, emitiendo opiniones acerca de lugares, productos o situaciones vividas. Estas opiniones emitidas a través de las redes sociales con el fin de publicar los sentimientos experimentados son de vital importancia a la hora de rescatar información para las empresas, con el fin apoyar en las decisiones comerciales.

En esta investigación se dieron a conocer fundamentos y conceptos básicos para abordar la problemática planteada al comienzo, dando a conocer diversos estilos y métodos para tratar de llegar a una solución óptima.

Se incorporó el análisis de sentimiento como método para encontrar la polaridad de las opiniones generadas, la caracterización de los datos con el fin de clasificar o etiquetar según características determinadas por la frecuencia de palabras. Con esto se dio pie a investigar sobre el aprendizaje automático incluyendo métodos como el K-NN y K- Means.

Se dio paso a la experimentación, dando pie a la generación de datos estadísticos, logrando comprar diversos ambientes y contextos entregados a los algoritmos. Para la primera experimentación (Twitter), se abordaron 4 tipos de dimensiones comparativas Ponderación de datos, Lematización de oraciones, N-gramas y 2 algoritmos de aprendizaje automático.

La primera experimentación sirvió de base para orientar el enfoque de la segunda (Reuters), esto con el fin de revalidar resultados obtenidos a través de los tweeter. Como se pudo apresar se logró dar énfasis en la ponderación TF-RFL y al método planteado en este estudio, el filtrado de datos. Luego se procedió a entregar tanto los resultados como sus respectivas conclusiones, las cuales son un punto de comparación a la primera experimentación.

Una vez analizadas ambas conclusiones, se puede desprender lo siguiente:

- La incorporación de la ponderación TF-RFL aumenta el índice de aciertos generados por los algoritmos cuando el conjunto de datos entregados para la clasificación posee más de dos categorías, reduciendo su asertividad cuando no.
- El filtrado de los datos a través de la ponderación TF-RFL resulto en una gran mejora de los resultados, cuando los conjuntos de datos poseen instancias similares y de más de dos categorías, por lo que sería método a considerar en la clasificación de sentimiento, por las propiedades que estos datos suelen tener.
- El proceso de lematización no dio resultados esperados para los algoritmos, llegando a considerarse un aspecto irrelevante en los resultados.
- Se pudo hacer una directa comparativa entre ambos algoritmos, de lo que se puede concluir que K-NN posee una gran ventaja en cuanto a aciertos en el ámbito del análisis de sentimiento bajo las circunstancias citadas con anterioridad. A su vez, el algoritmo K-Means no entrego los datos esperados, estos fueron muy bajos en comparación a su par.

- Se puede observar en ambas experimentaciones, que los resultados que brindan los algoritmos al aumentar el porcentaje de entrenamiento son de bajo impacto, lo que da a entender la eficiencia del algoritmo al momento de disminuir este conjunto, no necesitando de una gran cantidad de datos para obtener resultados positivos.

Como conclusión general del estudio, el algoritmo K-NN puede ser una herramienta muy potente para la clasificación automática, aun siendo de una baja complejidad, inclusive el algoritmo permite una gran movilidad a la hora de incorporarle diversos contexto de pruebas como lo fue en este estudio.

17 Bibliografía

[1] ITU NEWS, Los más destacado de El mundo en 2013: datos y cifras relativos las TIC. Información disponible vía web en https://itunews.itu.int/es/3781-Lo-mas-destacado-de-El-mundo-en-2013-datos-y-cifras-relativos-a-las-TIC.note.aspx. Revisada por última vez el 22 de octubre de 2015.

[2] Empresa Twitter, Sitio referido a empresas. Información disponible vía web enhttps://about.twitter.com/company . Revisada por última vez el 22 de octubre de 2015.

[3] José Carlos Cortizo Pérez, Minería de Opiniones. Información disponible vía web en http://www.brainsins.com/es/blog/mineria-opiniones/3555. Revisada por última vez el 22 de octubre de 2015.

[4] Mejova Y. (2010), Sentiment Analysis: AnOverview.

[5] Felipe Ignacio Oliva Valdebenito (2014), Minería de Opinión y Análisis de Sentimiento.

[6] Liu, B. (2006). Web Data Mining, Chapter Opinion Mining. Springer.

[7] Rosa María Coyotl Morales (2007), Clasificación automática de Textos considerando el Estilo de Redacción.

[8] Jacinto Dávila (2006), Lógica Práctica y Aprendizaje Computacional

[9] Cristina García Cambronero, Irene Gómez Moreno (2006), Algoritmo de Aprendizaje: KNN y Kmeans.

[10] Rodrigo Alfaro, Héctor Allende (2010), Text Representation in Multi-label Classification: Two New Input Representations

[11] Teledet, Estimación de la exactitud de una clasificación. Información disponible vía web en http://www.teledet.com.uy/tutorial-imagenes-satelitales/clasificacion-matriz-confusion.htm. Revisada por última vez el 17 de diciembre de 2015.

[12] Oldemar Rodríguez, Aprendizaje Supervisado. Información disponible vía web en http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentaci%C3%B3n_-_KNN.20085205.pdf. Revisada por última vez el 17 de diciembre de 2015.

[13] Ramón Hermoso y MatteoVesirani, Aprendizaje inductivo. Información disponible vía web en http://www.ia.urjc.es/cms/sites/default/files/userfiles/file/ia4/2011/IA4_%5BArboles%5D%281%29.pdf.Revisada por última vez el 17 de diciembre de 2015.

[14] Jdiez, Juanjo, Sistemas Inteligentes: Aprendizaje no Supervisado. Información disponible vía web en http://www.aic.uniovi.es/ssii/ssii-t12-aprendizajenosupervisado.pdf. Revisada por última vez el 17 de diciembre de 2015.

[15] Carles M Cuadras (1989), Distancias Estadísticas, Universidad de Barcelona.

[16] Vitutor, Modulo de un vector. Información disponible vía web enhttp://www.vitutor.com/geo/vec/a_4.html. Revisada por última vez el 17 de diciembre de 2015.

[17] Aurea Grané, Escalado Multidimensional. Información disponible vía web enhttp://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_Coorp_reducido.pdf. Revisada por última vez el 17 de diciembre de 2015.

[18] Criterios de similitud. Similitud, divergencia y distancia. Información disponible vía web enhttps://www.uv.es/ceaces/multivari/cluster/criterios_de_similitud.htm. Revisada por última vez el 17 de diciembre de 2015.

[19] Análisis de Cluster y Multidimensional Scaling. Información disponible vía web enhttp://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema5am.pdf. Revisada por última vez el 17 de diciembre de 2015.

[20] Miguel Cárdenas-Montes, Medidas de Distancia. Información disponible vía web enhttp://wwwae.ciemat.es/~cardenas/docs/lessons/MedidasdeDistancia.pdf. Revisada por última vez el 17 de diciembre de 2015.

[21] Jojjoa-tecnología marketing y crm, Definición de clustering. Información disponible vía web enhttps://sites.google.com/site/jojooa/inteligencia-artificial/definicion-de-clustering-que-es-el-clustering. Revisada por última vez el 18 de diciembre de 2015.

[22] Abdelmalik Moujahid y Pedro Larrañaga. Clasificadores K-NN. Información disponible vía web enhttp://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t9knn.pdf. Revisada por última vez el 18 de diciembre de 2015.

[23] Introducción al análisis cluster. Información disponible vía web enhttps://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm. Revisada por última vez el 18 de diciembre de 2015.

[24] Jorge Alonso Bedoya Puerta, Aplicación de distancias entre términos para datos planos y jerárquicos. Información disponible vía web enhttp://users.dsic.upv.es/~flip/papErs/TFM-JorgeBedoya.pdf.

[25] Sung-Hyuk Cha y Sargur N. On measuring the distance between histograms (2002).

[26] Sung-Hyuk Cha. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions (2007).

[27] Online Language Dictionaries. Información disponible vía web en http://www.wordreference.com/definicion/corpus. Revisada por última vez el 21 de abril del 2016.

[28] Weka 3: Data Mining Software in Java. Información disponible vía web en http://www.cs.waikato.ac.nz/ml/weka. Revisada por última vez el 21 de abril del 2016.

[29] Download Reuters and Ohsumed. Información disponible vía web en https://www.mat.unical.it/OlexSuite/ Datasets/SampleDataSets-download.htm. Revisada por última vez el 14 de marzo de 2017.

[30] Matriz de confusión. Información disponible vía web enhttp://brenocon.com/confusion_matrix_diagrams.pdf. Revisada por última vez el 14 de marzo de 2017.

Anexos

Anexo A: Resultados K-NN

Anexo A.1: K-NN Sin lematizar

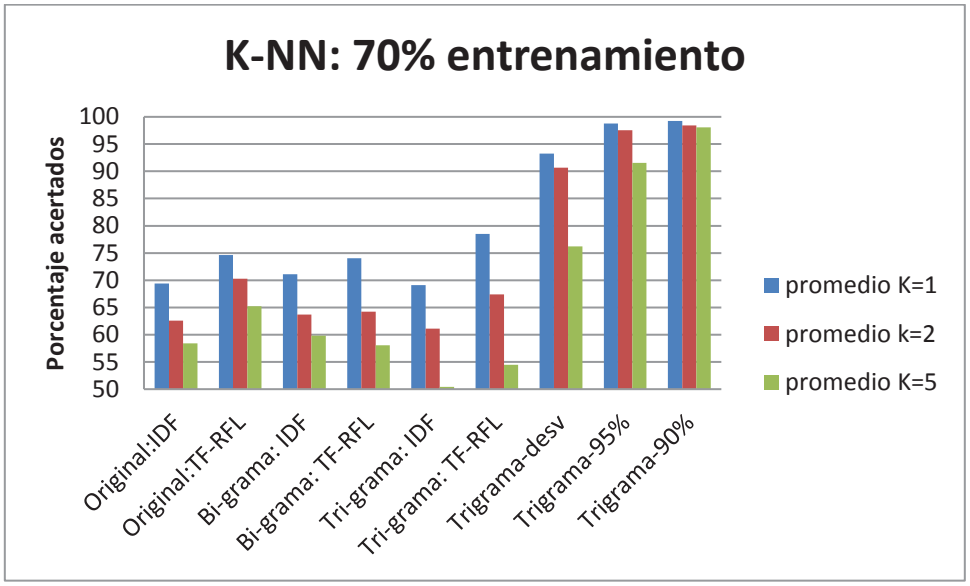

Figura 0.1 70% Sin Lematizar

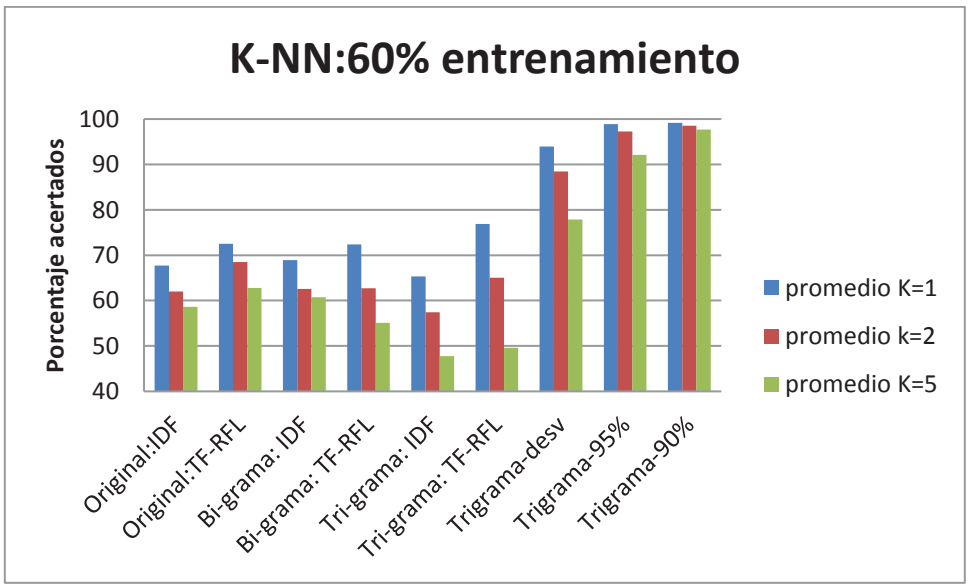

Figura 0.2 60% Sin Lematizar

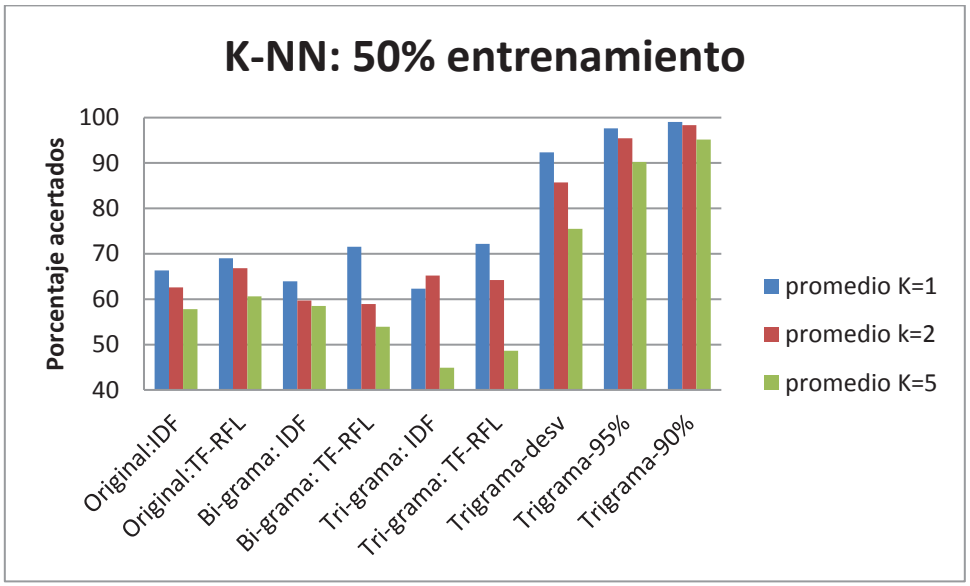

Figura 0.3 50% Sin Lematizar

Anexo A.2 K-NN Lematizado

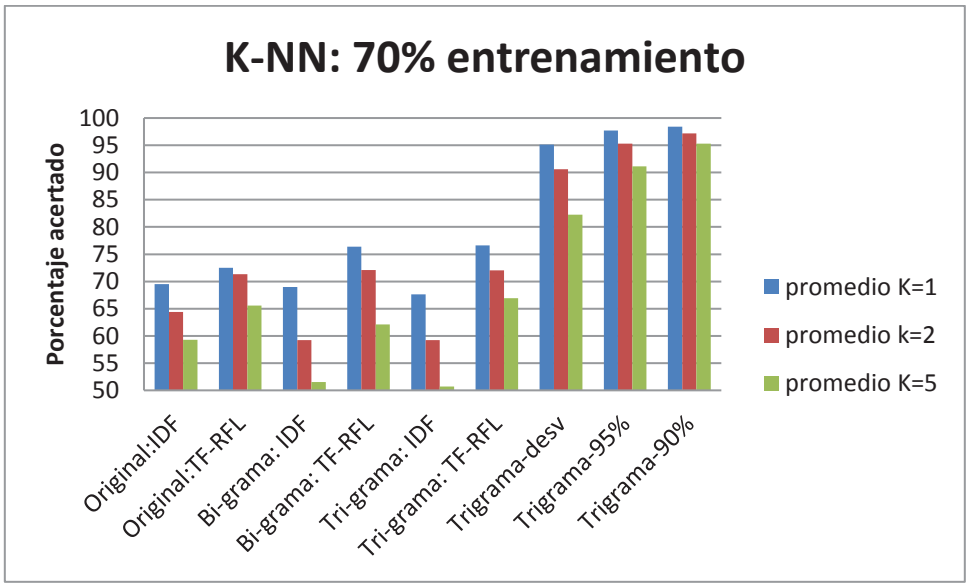

Figura 0.4 70% Lematizado

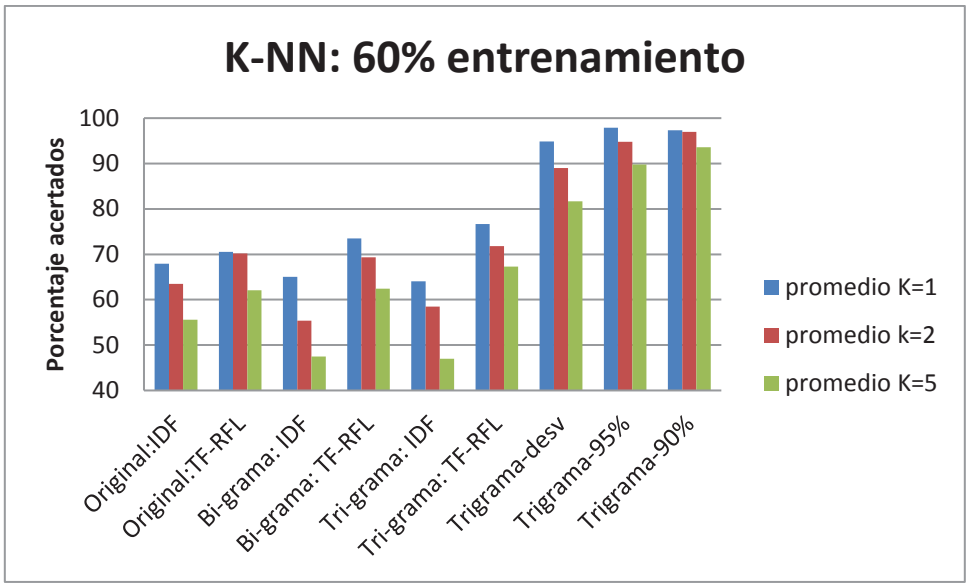

Figura 0.5 60% Lematizado

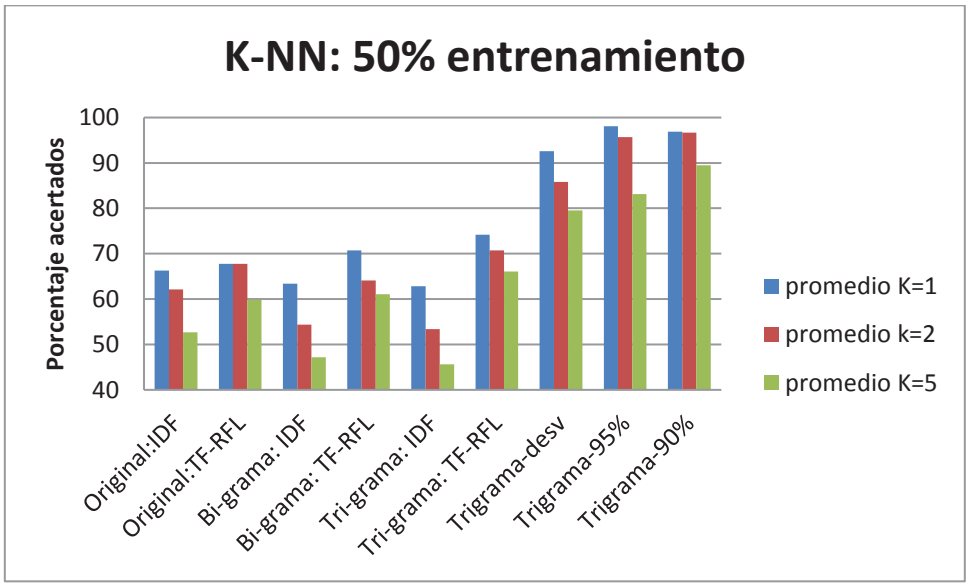

Figura 0.6 50% Lematizado

Anexo B: K-Means

Tabla 0.1 Datos resumen K-Means

	Sin Lematizar	Lematizado
Original IDF	34,7	40,5
Original TF-RFL	41,86	43,6
Bigrama-IDF	36,08	36,2
Bigrama- TF-RFL	42,2	34,97
Trigrama-IDF	33,7	34,9
Trigrama-TF_RFL	37,2	33,9
Desv-RFL-Bigrama	36,022	47,76
Desv-RFL-TRI	47,31	42,96
90-TF-RFL-Trigrama	56	53,81
95-TF-RFL-Trigrama	42	47

Anexo C: Tablas datos K-NN Sin Lematizar

Tabla 0.1 Original: IDF

Original: IDF	K=1	K=2	K=5
95	66,6	54,66	60
90	70,66	64,66	60
80	71,6	64,6	61
70	69,4	62,6	58,4
60	67,75	62	58,6
50	66,3	62,6	57,8

Tabla 0.2 Original: TF-RFL

Original: TF-RFL	K=1	K=2	K=5
95	74	74	67
90	74,6	72	65,3
80	75	71	67,5
70	74,6	70,3	65,2
60	72,5	68,5	62,75
50	69	66,8	60,6

Tabla 0.3 Bigrama: IDF

Bigrama: IDF	K=1	K=2	K=5
95	74	60,6	61,3
90	70,6	62,3	62,6
80	70,8	63,8	63,8
70	71,1	63,7	59,8
60	68,9	62,6	60,75
50	63,9	59,7	58,5

Tabla 0.4 Bigrama: TF-RFL

Bigrama: TF-RFL	K=1	K=2	K=5
95	72,1	62,5	59,8
90	76,19	64,6	63,2
80	73,6	63,4	60,2
70	74,03	64,2	58,04
60	72,4	62,7	55,1
50	71,54	58,95	53,9

Tabla 0.5 Trigrama: IDF

Trigrama: IDF	K=1	K=2	K=5
95	74,6	62,6	54
90	69,6	60,6	51
80	69,6	62	54,1
70	69,1	61,1	50,4
60	65,3	57,4	47,75
50	62,3	65,2	44,9

Tabla 0.6 Trigrama: TF-RFL

Trigrama: TF-RFL	K=1	K=2	K=5
95	83,5	71,6	52,9
90	86,9	76,5	54,2
80	82,7	72,6	54,8
70	78,5	67,4	54,5
60	76,9	65,02	49,6
50	72,2	64,2	48,6

Tabla 0.7 Desv-TF-RFL-Original

Desv-TF-RFL-Original	K=1	K=2	K=5
95	90,72	87,6	87,6
90	90,76	89,2	89,74
80	90,76	87,9	87,17
70	90,23	88,1	84,4
60	87,41	85,2	84,08
50	86,55	85,1	82,85

Tabla 0.8 Desv-TF-RFL-Bigrama

Desv-TF-RFL-Bigrama	K=1	K=2	K=5
95	95,65	93,47	80,4
90	94,53	91,8	75,9
80	95,09	89,9	79,2
70	94	88,72	80,72
60	92,37	86,6	77,79
50	90,94	84,84	77,75

Tabla 0.9 Desv- TF-RFL- Trigrama

Desv- TF-RFL- Trigrama	K=1	K=2	K=5
95	97,4	94,8	85,71
90	97,4	94,15	83,76
80	94,4	89,28	78,89
70	93,24	90,69	76,19
60	93,94	88,47	77,9
50	92,33	85,71	75,5

Tabla 0.10 95%-TF-RFL- Trigrama

95%-TF-RFL- Trigrama	K=1	K=2	K=5
95	98,1	98,18	90,9
90	99,09	99,09	92,72
80	99,09	96,83	92,3
70	98,79	97,55	91,54
60	98,86	97,28	92,08
50	97,64	95,47	90,21

Tabla 0.11 90% -TF-RFL- Trigrama

90% -TF-RFL- Trigrama	K=1	K=2	K=5
95	95,3	95,34	95,34
90	97,7	97,7	97,76
80	98,8	98,2	98,26
70	99,23	98,4	98,07
60	99,13	98,5	97,68
50	99,07	98,3	95,15

Anexo D: Tablas datos K-NN Lematizado

Tabla 0.1 Original: IDF

Original: IDF	K=1	K=2	K=5
95	70	70	62
90	72,6	69,6	60,6
80	70,3	66,5	60
70	69,5	64,4	59,3
60	67,9	63,5	55,6
50	66,3	62,1	52,7

Tabla 0.2 Original: TF-RFL

Original: TF-RFL	K=1	K=2	K=5
95	75,3	73,3	66
90	74,6	72	69
80	73,3	71,3	65,6
70	72,5	71,3	65,6
60	70,5	70,2	62,08
50	67,8	67,8	59,9

Tabla 0.3 Bigrama: IDF

Bigrama: IDF	K=1	K=2	K=5
95	70	60	49,3
90	68,6	60,3	53,3
80	69,1	58,6	52
70	69	59,2	51,5
60	65	55,4	47,5
50	63,4	54,4	47,2

Tabla 0.4 Bigrama: TF-RFL

Bigrama: TF-RFL	K=1	K=2	K=5
95	79,5	72,78	65,98
90	79,1	72,69	60
80	77,3	73,2	63
70	76,4	72,1	62,1
60	73,5	69,3	62,4
50	70,75	64,1	61,1

Tabla 0.5 Trigrama: IDF

Trigrama: IDF	K=1	K=2	K=5
95	70,6	60,6	46,6
90	69,3	60	50
80	68,6	60,1	51,5
70	67,6	59,2	50,7
60	64,08	58,5	47
50	62,8	53,4	45,6

Tabla 0.6 Trigrama: TF-RFL

Trigrama: TF-RFL	K=1	K=2	K=5
95	83,9	76,64	74,4
90	82,1	77,7	70,8
80	77,7	73,7	68,6
70	76,6	72,05	66,91
60	76,7	71,8	67,3
50	74,2	70,7	66,05

Tabla 0.7 Desv-TF-RFL-Original

Desv-TF-RFL-Original	K=1	K=2	K=5
95	88,77	88,77	82,65
90	88,26	88,26	83,67
80	84,54	89,54	77,5
70	88,77	88,77	79,93
60	88,26	88,26	80,9
50	87,24	87,24	80,91

Tabla 0.8 Desv-TF-RFL-Bigrama

Desv-TF-RFL-Bigrama	K=1	K=2	K=5
95	91,2	87,91	79,12
90	93,95	92,3	81,42
80	94,79	90,41	82,46
70	92,51	89,05	81,02
60	92,6	89,17	78,08
50	91,01	85,1	76,34

Tabla 0.9 Desv- TF-RFL- Trigrama

Desv- TF-RFL- Trigrama	K=1	K=2	K=5
95	100	95	78,75
90	95,625	92,5	80,62
80	95,61	89,6	81,19
70	95,15	90,6	82,25
60	94,83	89,04	81,69
50	92,6	85,83	79,57

Tabla 0.10 95%-TF-RFL- Trigrama

95%-TF-RFL- Trigrama	K=1	K=2	K=5
95	100	100	88,5
90	97,5	96,6	90,9
80	98,7	96,6	92,1
70	97,7	95,3	91,1
60	97,9	94,8	89,8
50	98,1	95,7	83,1

Tabla 0.11 90% -TF-RFL- Trigrama

90% -TF-RFL- Trigrama	K=1	K=2	K=5
95	100	100	97,6
90	100	98,82	96,4
80	99,41	98,24	94,7
70	98,43	97,2	95,31
60	97,3	97	93,56
50	96,9	96,7	89,46

Anexo E: REUTERS

Tabla 0.1 Reuters 50%

50%	Ponderación	Promedio
K=30	IDF	83,90
	TRFL	77,46
	TRFL 5%	69,76
	TRFL 10%	67,32
K=10	IDF	86,99
	TRFL	85,13
	TRFL 5%	79,36
	TRFL 10%	77,04
K=5	IDF	88,50
	TRFL	86,21
	TRFL 5%	80,75
	TRFL 10%	78,69
K=1	IDF	89,27
	TRFL	87,32
	TRFL 5%	86,04
	TRFL 10%	83,22

Tabla 0.2 Reuters 60%

60%	Ponderación	Promedio
K=30	IDF	83,98
	TRFL	78,19
	TRFL 5%	70,73
	TRFL 10%	68,20
K=10	IDF	86,78
	TRFL	86,39
	TRFL 5%	80,03
	TRFL 10%	76,38
K=5	IDF	88,21
	TRFL	87,67
	TRFL 5%	81,57
	TRFL 10%	80,24
K=1	IDF	89,36
	TRFL	88,49
	TRFL 5%	86,50
	TRFL 10%	84,09

Tabla 0.3 Reuters 70%

70%	Ponderación	Promedio
K=30	IDF	84,47
	TRFL	79,08
	TRFL 5%	71,14
	TRFL 10%	68,39
K=10	IDF	87,27
	TRFL	87,00
	TRFL 5%	80,95
	TRFL 10%	76,92
K=5	IDF	88,40
	TRFL	87,36
	TRFL 5%	82,04
	TRFL 10%	80,83
K=1	IDF	89,66
	TRFL	88,75
	TRFL 5%	87,03
	TRFL 10%	84,77

Anexo E.1: Gráficos comparativos Reuters

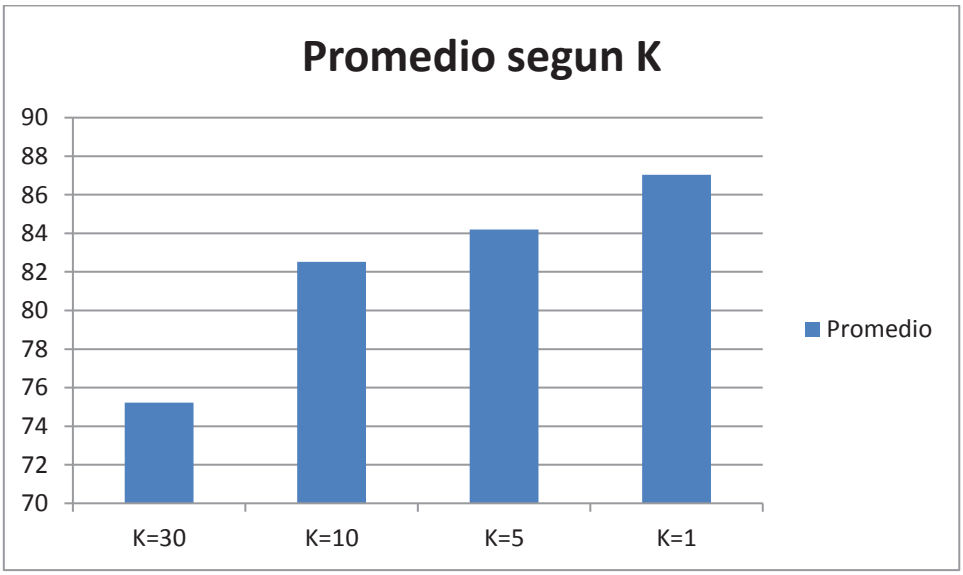

Figura 0.1 Promedio Reuters según K

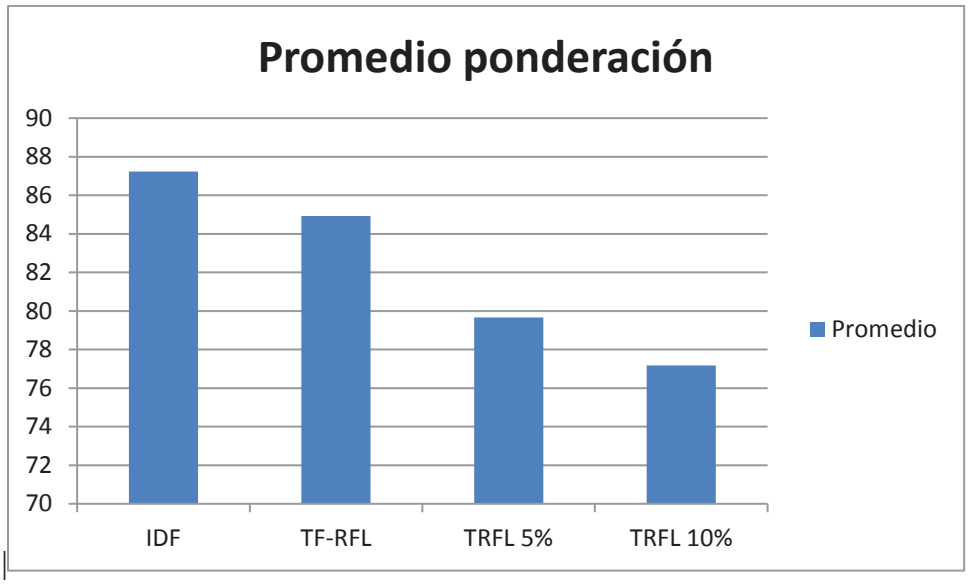

Figura 0.8 Promedio Reuters según Ponderación

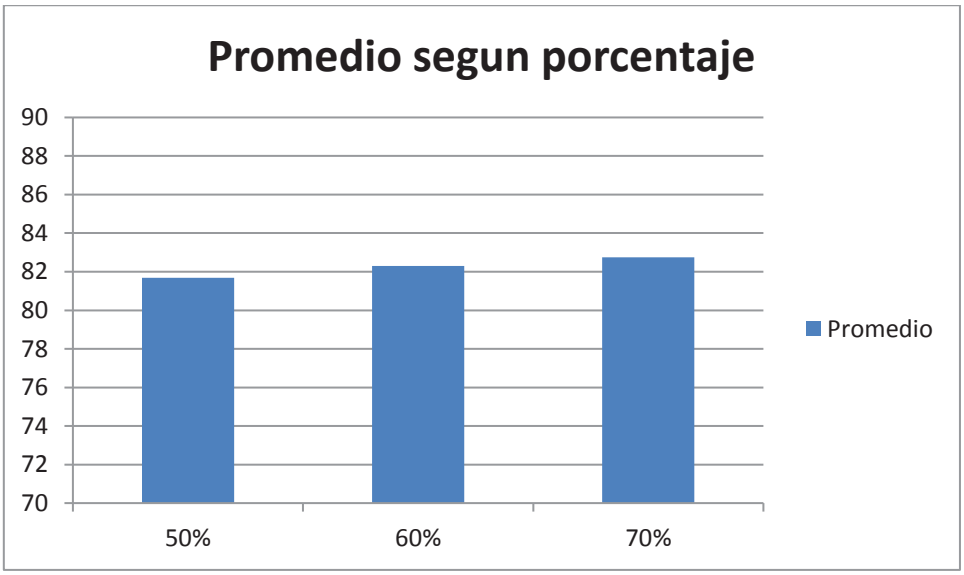

Figura 0.92 Promedio Reuters según Porcentaje