PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

“Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente”

Sebastian Ariel Pérez Vera

Profesor Guía: **Rodrigo Alfaro Arancibia**

Profesor Co-referente: **Héctor Allende Cid**

Carrera: **Ingeniería Civil en Informática**

Abril del 2017

Resumen

En los últimos años la clasificación de documentos basada en la opinión (conocida en inglés bajo los nombres de sentiment classification, sentiment analysis u opinion mining) ha sido objeto de un creciente interés por parte de la comunidad de investigadores del procesamiento del lenguaje natural.

El creciente interés por el procesamiento automático de las opiniones contenidas en documentos de texto, es en parte consecuencia del aumento exponencial de contenidos generados por usuarios en la Web 2.0, y por el interés, entre otros, de empresas y administraciones públicas en analizar, filtrar o detectar automáticamente las opiniones vertidas por sus clientes o ciudadanos.

El presente trabajo ha tenido como objetivo la implementación de un sistema de clasificación automática de textos de opiniones, concretamente de las opiniones realizadas por los clientes de una compañía eléctrica en su área de atención específicamente en su cuenta Twitter que tiene destinada para esta finalidad. Los documentos (Tweets generados por los usuarios) fueron clasificados según categorías establecidas (Pregunta, Sugerencia, Información Usuarios, Otras e Información Empresa), aplicando diferentes técnicas como Arboles de decisión, Maquina de Soporte Vectorial, Algoritmo de Nearest Neighbors y algunas variantes de Naive Bayes. Con la implementación de este sistema se ha buscado facilitar la clasificación manual de estas opiniones y permitir una rápida respuesta a los documentos que lo requieran.

Los resultados obtenidos tras las distintas pruebas que se realizaron han demostrado ser bastantes satisfactorios lográndose los mejores resultados con el método de Arboles de Decisión, en los cuales se obtuvo un promedio tras 10 rondas de prueba de 97% de Accuracy utilizando la representación TF-RFL, seguido de la representación TF-RFL con el método KNN con la cual se ha logrado un valor de 91,4% Accuracy.

Con la realización del proyecto, se comprobó y analizo también las dificultades encontradas en la implementación de un sistema de clasificación automática donde la naturaleza de los textos es de opinión.

Palabras Claves: predicción, resultados, clasificadores, sentiment analysis u opinión mining

Abstract

In recent years, opinion-based classification of documents (sentiment classification, sentiment analysis, or opinion mining) has been the subject of growing interest on the part of the community of natural language processing researchers.

The growing interest in automatic processing of the opinions contained in text documents is partly a result of the exponential increase in content generated by users in Web 2.0, and the interest, among others, in companies and public administrations in analyzing, filtering Or automatically detect the opinions expressed by their clients or citizens.

The objective of the present work was to implement a system of automatic classification of texts of opinions, specifically the opinions made by customers of an electric company in their area of attention specifically in their Twitter account that is intended for this purpose. The documents (Tweets generated by the users) were classified according to established categories (Question, Suggestion, Information Users, Others and Company Information), applying different techniques such as Decision Trees, Support Vector Machines, Nearest Neighbors Algorithm and some Naive Bayes variants. The implementation of this system has sought to facilitate the manual classification of these opinions and allow a quick response to the documents that require it.

The results obtained after the different tests that have been performed have proved to be quite satisfactory, obtaining the best results with the Decision Trees method in which an average of 10 test rounds of 97% of Accuracy was obtained using the TF-RFL representation, Followed by the TF-RFL representation with the KNN method with which a value of 91.4% Accuracy has been achieved.

With the realization of the project, it was also verified and analyzed the difficulties encountered in the implementation of an automatic classification system where the nature of the texts is of opinion.

Keywords : prediction , results, classifiers , sentiment analysis or opinion mining

Índice

Índice de Figuras

Índice de Tablas

1.- Introducción

1.1.- Introducción

Desde el comienzo de nuestra historia los humanos han buscado la forma de comunicarse, característica que nos hace únicos, al poder pensar de forma racional y poder interactuar con otros, de aquí nace la escritura. Las personas han plasmado conocimientos, sentimientos y emociones para crear documentos, poesía e incluso verdaderas obras maestras, pero todas ellas, indiferente lo que expresen, contienen un denominador común, información.

La cual es sumamente importante para la supervivencia del ser humano actual, ya que con el paso del tiempo y la incorporación de la tecnología, es que todo el conjunto de documentos han ido creando una gran biblioteca de información digital. Para que estos documentos digitales sean de fácil acceso se han tenido que organizar de tal manera que se permita su recuperación y análisis por medios automáticos. Sin embargo, el problema es complejo y es necesaria la continua investigación en la búsqueda de métodos y representaciones apropiadas. Es por ello que se han creado diversas líneas de investigación para el tratamiento automático de textos, entre las que se encuentran: la recuperación de información, la extracción de información, la búsqueda de respuestas y la clasificación de textos entre otras.

Hoy en día el uso frecuente de internet y de las redes sociales como Facebook, Twitter entre otras ha llevado a que dicha información crezca de una manera abrumadora, ya que, las personas plasman sus opiniones, comentarios, sentimientos, anhelos y frustraciones de sí mismos o de su entorno. Con esto nace una pregunta ¿Qué es lo que las personas piensan? y esta pregunta puede llegar a ser respondida en la actualidad mediante la extracción, clasificación y análisis de lo que la gente escribe en el mundo digital, por ello empresas han apuntado a esta investigación como proceso clave en la toma de decisiones, el poder saber lo que la gente opina y "siente" respecto a un determinado producto o campaña es indispensable al momento de actuar para tomar las medidas correspondientes.

En el presente se abordara la clasificación de documentos de una compañía eléctrica en categorías definidas para luego ver si es posible dar respuestas automáticas a ciertos documentos que sean de contenido común. Con el fin de facilitar la toma de decisiones y el manejo del gran volumen de información que se recibe en la actualidad.

1.2.- Objetivos

1.2.1.- Objetivo General

El objetivo que se quiere lograr con el presente trabajo es la clasificación automática de las opiniones vertidas en los canales de comunicación de las tiendas del Retail u otras compañías a través de las Redes Sociales (Twitter), clasificándolo en categorías, mediante la aplicación de diversas técnicas de clasificación. Para poder inferir respuestas automáticas a estos comentarios para así agilizar las tareas que son realizadas de forma manuales.

1.2.2.- Objetivos Específicos

- Investigar distintas técnicas de clasificación de textos.
- Interpretar el funcionamiento de las técnicas vistas.
- Ser capaz de extraer la información relevante contenida en los Tweet.
- Desarrollar un sistema que permita clasificar las opiniones.
- Crear set de respuestas tipo a preguntas frecuentes.

1.3.- Estructura del Documento

Introducción: Contendrá la introducción al tema abordado junto con el planteamiento de los objetivos que se quieren lograr con la realización del presente trabajo.

Definición del Problema: Contendrá la definición de la problemática que abordara el problema, la cual permitirá entender de una mejor manera los objetivos planteados.

Marco Teórico: Este capítulo se refiere al marco teórico que hay detrás de la problemática que se ha visto en el apartado anterior, identificando cuales son las distintas técnicas que permitirán llevar a cabo los objetivos planteados.

Twitter y sus Características: Este capítulo pretende ser una guía para entender un poco más de que es Twitter y como están estructurados sus mensajes.

Arquitectura del Sistema: En este capítulo se describirán todos los pasos necesarias para realizar la clasificación de los datos, partiendo por la definición de la categorías y terminando con la representación de los datos en forma vectorial utilizando diferentes métodos de asignación de pesos.

Experimentación de Clasificadores: En este capítulo se mostraran los resultados obtenidos tras realizar los diferentes experimentos.

Respuestas Automatizadas: Este capítulo contendrá una breve reseña de lo que son las respuestas automáticas, la identificación de las respuestas que se han considerado que pueden ser automatizadas para finalmente mostrar las métricas correspondientes a la clasificación de estas.

Conclusiones y Trabajos Futuros: Finalmente se presentaran las distintas conclusiones obtenidas con la realización del presente trabajo como también los trabajos futuros que se realizaran.

2.- Definición del Problema

2.1.- Definición del Problema

La atención al cliente ha evolucionado en los últimos años permitiendo a las empresas y marcas estar más comunicados con sus clientes, de una manera cercana, transparente e inmediata. La atención al cliente a través de redes sociales permite ahora más que nunca a las organizaciones demostrar su promesa de estar cerca del cliente, de resolver dudas y de generar valor.

Atender al cliente a través de las redes sociales no es únicamente cuestión de encontrar una herramienta tecnológica que nos resuelva en 3 minutos las dudas de los clientes. El reto está en escuchar plenamente lo que los clientes quieren y cómo desean ser atendidos por todos los canales que esto implique.

El estudio titulado “Social Media Benchmark Study 2013” [1] nos dice que el 67% de los consumidores utilizan los perfiles en redes sociales de una empresa para obtener información sobre los servicios y que el 43% de los jóvenes de entre 18 y 29 años, prefiere interactuar con las marcas a través de las redes sociales, que utilizar otro tipo de comunicación. Es por esto que nace la problemática de cómo interpretar esta enorme cantidad de opiniones que los distintos clientes y usuarios de marcas y empresas están realizando en las redes sociales (conocida en inglés bajo los nombres de sentiment classification, sentiment analysis u opinion mining).

Entonces una vez mencionado lo anterior, resulta de suma importancia encontrar herramientas que permitan manejar ese gran volumen de información para ser tratado y analizado de forma pertinente que ayude a encontrar las respuestas adecuadas para el correcto funcionamiento de las entidades que utilizan las redes sociales como fuentes de atención.

3.- Marco Teórico

3.1.- Planteamiento Inicial

Una gran parte de los contenidos generados por los usuarios en forma de textos escritos en lenguaje natural tiene un carácter subjetivo: los reviews de productos o servicios, las entradas de algunos blogs o los comentarios hechos por otros usuarios, los hilos de discusión en los foros públicos (acerca de política, cultura, economía o cualquier otro tema), los mensajes escritos en servicios de microblogging como Twitter o en redes sociales como Facebook. Todos estos son textos en los que los internautas pueden expresar sus opiniones, puntos de vista y estados de ánimo. Por otro lado, son evidentes las implicaciones prácticas de este tipo de contenidos para las compañías o administraciones públicas; por ejemplo, hoy día muchas compañías disponen de personal que se dedica a monitorizar las opiniones aparecidas en Internet acerca de los productos y servicios de la misma, de manera que puedan interferir activamente evitando la propagación de estas opiniones y sus posibles consecuencias negativas de cara a las ventas o a la imagen de la marca. Pero independientemente del objetivo buscado, el número de contenidos individuales que deberían ser considerados es de tal magnitud que se hacen imprescindibles ciertos mecanismos de procesamiento automático de los mismos. Así se abre una nueva subdisciplina conocida como Análisis del Sentimiento o Minería de Opiniones que viene a ocuparse del procesamiento computacional de este tipo de contenidos subjetivos.

Así, el análisis del sentimiento (sentiment analysis) o minería de opiniones (opinion mining) abarca aquellas tareas relacionadas con el tratamiento computacional de las opiniones, los sentimientos y otros fenómenos subjetivos del lenguaje natural. Algunas de estas tareas son la clasificación de documentos de opinión según el carácter positivo o negativo de las opiniones, la extracción de representaciones estructuradas de opiniones, o el resumen de textos de opinión, entre otras. [2]

La clasificación automática de textos es por lo general un proceso supervisado Esto significa que requiere de un conjunto de documentos previamente clasificados por expertos humanos que funcionan como entrenamiento para el sistema. Así, un conjunto de documentos clasificado por un humano en cierta categoría sirven para que el clasificador automático genere una clasificación propia frente a un documento desconocido. El desempeño del clasificador automático dependerá de qué tan similar sea esta clasificación respecto a la humana, lo que se evalúa con matrices de confusión y otras métricas [3].

Como se mencionó anteriormente la clasificación automática de texto consiste en un conjunto de algoritmos, técnicas y sistemas capaces de asignar un documento a una o varias

categorías o grupos de documentos, construidos según su afinidad temática. Para ello, véase Figura 1, se emplean técnicas de Aprendizaje Automático (ML, Machine Learning) y de Procesamiento de Lenguaje Natural (NLP, Natural Language Processing).

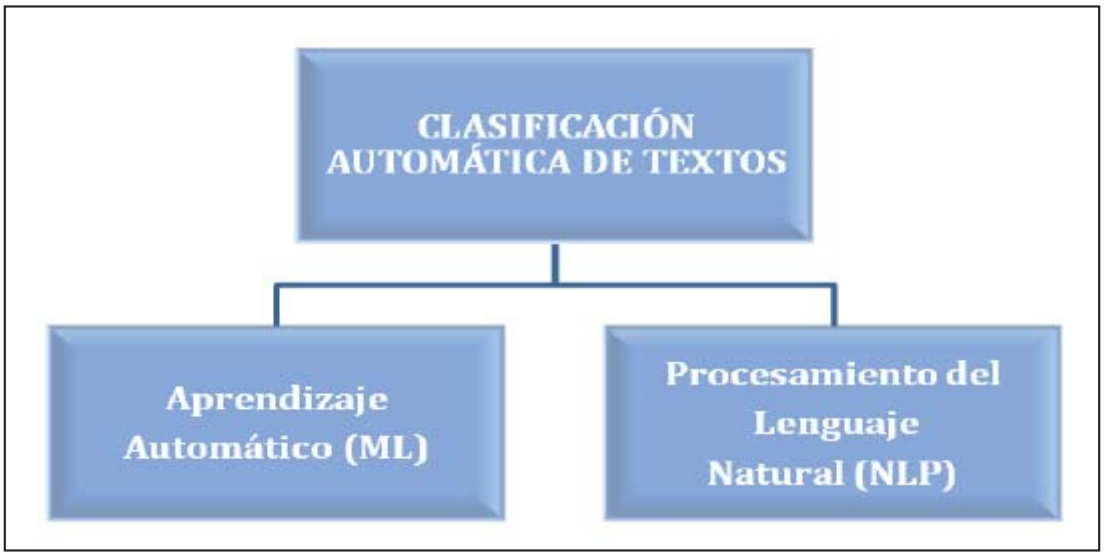


Figura 1.- Clasificación automática de texto

El Procesamiento Natural del Lenguaje (PNL) estudia los problemas inherentes al procesamiento y manipulación de lenguajes naturales, haciendo uso de ordenadores. Pretende adquirir conocimiento sobre el modo en que los humanos entienden y utilizan el lenguaje, de tal forma que se pueda llevar a cabo el desarrollo de herramientas y técnicas para conseguir que los ordenadores puedan entenderlo y manipularlo. Sus fundamentos residen en un conjunto muy amplio de disciplinas: ciencias de la información y los computadores, lingüística, matemáticas, ingeniería eléctrica y electrónica, inteligencia artificial y robótica, psicología, etc. Existe un gran número de aplicaciones donde el PNL resulta de gran utilidad (traducción máquina, procesamiento y resumen de textos escritos en lenguaje natural, interfaces de usuario, reconocimiento de voz, etc.).

Para el análisis sintáctico y morfológico de los textos. Extracción de Información, Clasificación Automática de Documentos, Agrupamiento Semántico, entre otras tareas se cubren utilizando técnicas de Machine Learning. Los modelos de Análisis de Sentimiento permiten reconocer la orientación o polaridad subjetiva de un texto, esto es, si están hablando bien o mal de aquello que se está opinando. Observemos la siguiente figura (Figura 2) para tener un entendimiento un poco más gráfico de los conceptos mencionados:

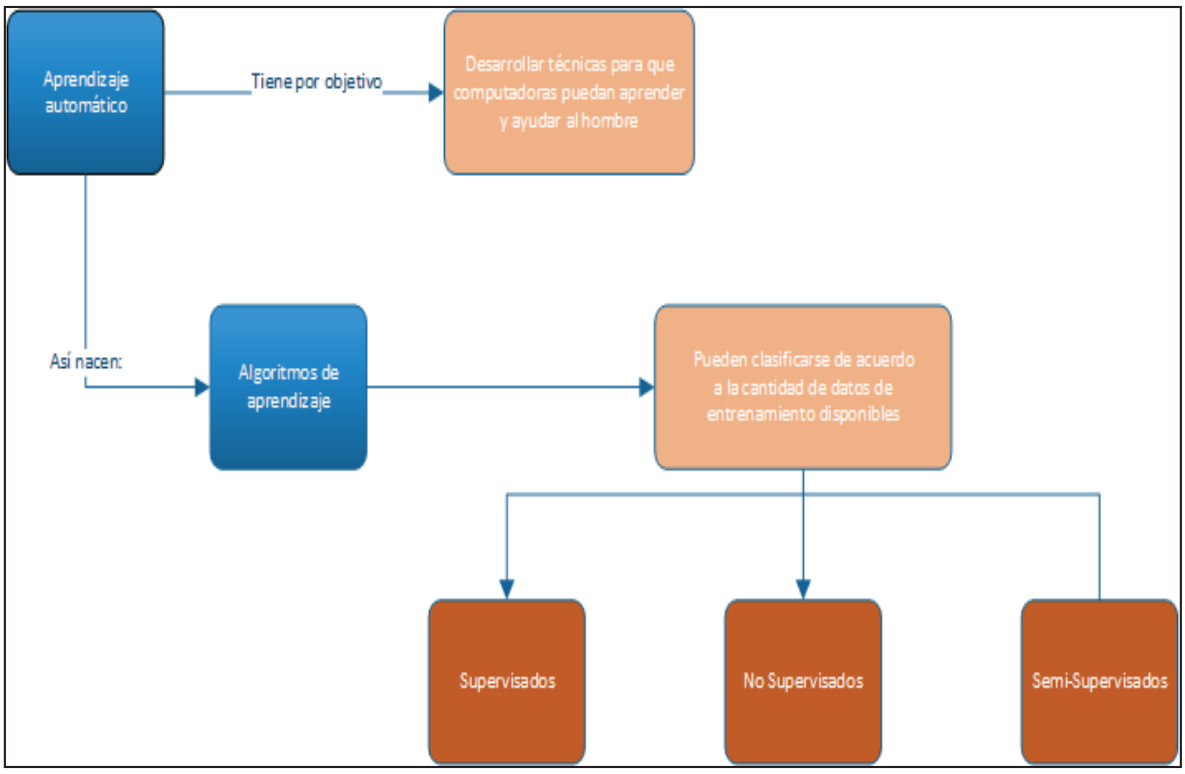


Figura 2.- Aprendizaje automático

3.2.- Tipos de Clasificadores

A demás de lo expuesto en la imagen anterior se puede especificar un clasificador de textos según distintos puntos de vista.

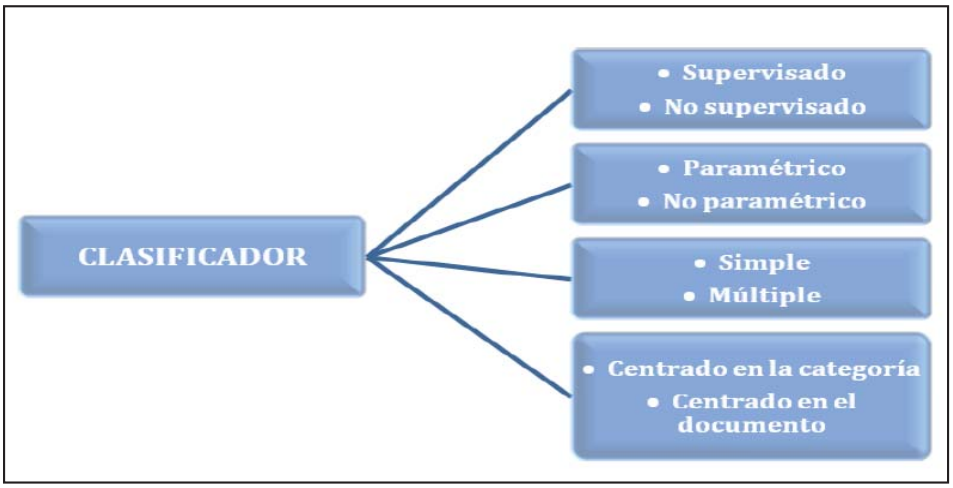


Figura 3.- Tipos de clasificadores

3.2.1.- Clasificación Supervisada y no Supervisada

3.2.1.1.- Clasificación Supervisada

Partiendo de una serie de categorías conceptuales prediseñadas a priori, se encarga de asignar cada documento a la categoría correspondiente. Requiere la elaboración manual o intelectual del conjunto de categorías. Además, es necesaria una fase de entrenamiento por parte del clasificador. El objetivo que se persigue en los clasificadores supervisados es el siguiente: elaborar un patrón representativo para cada una de las categorías entrenadas y aplicar alguna función que permita estimar la similitud entre el documento a clasificar y cada uno de estos patrones. Aquel patrón o patrones que presenten más concordancias con el documento indicarán la categoría o categorías a las que pertenece el mismo. El proceso de elaboración de los patrones necesita un conjunto de documentos previamente clasificados y se conoce como aprendizaje o entrenamiento.

3.2.1.2.- Clasificación No Supervisada

No existen categorías previas o cuadros de clasificación establecidos a priori. Los documentos se clasifican en función de su contenido, de forma automática, sin asistencia manual. Es una segmentación o agrupamiento automático, en inglés conocido como clustering. Una de las aplicaciones de este tipo de clasificación es la compresión de imágenes.

3.2.1.3.- Clasificación Semi-Supervisada

El aprendizaje semi-supervisado es un tipo de algoritmo de Machine Learning que hace uso de datos, tanto etiquetados como no etiquetados previamente como datos de entrenamientos, típicamente un pequeño número de datos etiquetados y una gran cantidad de datos no etiquetados. Este aprendizaje cae entre el aprendizaje no supervisado y el supervisado siendo una mezcla de los dos.

La obtención de datos previamente etiquetados presenta muchas veces un problema, ya que requiere un grupo de humanos calificados para la etiquetación manual de los ejemplos, muchos investigadores han descubierto que los datos no etiquetados junto con una pequeña cantidad de datos etiquetados aumenta considerablemente la efectividad del aprendizaje por lo que es una excelente opción cuando no se cuenta con muchos recursos.

3.2.2.- Clasificación Paramétrica y no Paramétrica

3.2.2.1.- Clasificación Paramétrica

En el entrenamiento de un clasificador paramétrico se emplea el set de entrenamiento para estimar o aprender los parámetros del modelo. El set de test que contiene documentos a clasificar se emplea para determinar la capacidad de generalización del clasificador [4].

3.2.2.2.- Clasificación No Paramétrica

Esta clasificación se subdivide en dos categorías:

La primera está basada en patrones y se obtiene una descripción de cada categoría en términos de un patrón, normalmente en forma de vector de términos con peso, como por ejemplo el clasificador Rocchio. La clasificación de los documentos se realiza en función de las similitudes existentes entre cada documento y los distintos patrones. [5]

La segunda categoría está basada en ejemplos y los documentos se clasifican según las similitudes que presenten con ejemplos del conjunto de entrenamiento. El clasificador más conocido es el del vecino más cercano (KNN, K-Nearest Neighbour).

3.2.3.- Clasificación Simple y Múltiple

3.2.3.1.- Clasificación Simple

Cada documento tiene una única categoría. Se trata de una clasificación donde las categorías no se solapan. Un caso especial es la clasificación binaria, donde cada documento pertenece a una categoría o a su complementaria.

3.2.3.2.- Clasificación Múltiple

Cada documento puede recibir un número variable de categorías. En este caso, las categorías sí se pueden solapar.

3.3.- Técnicas de Clasificación Automática de Textos

A continuación se mostraran algunas técnicas utilizadas en la clasificación automática de textos.

3.3.1.- Algoritmos Probabilísticos (Naive Bayes)

Se basan en la teoría probabilística, en especial en el teorema de Bayes, el cual permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero. El algoritmo más conocido, y también el más simple, es el denominado Naïve Bayes [6], que estima la probabilidad de que un documento pertenezca a una categoría. Dicha pertenencia depende de la posesión de una serie de características, de cada una de las cuales se conoce la probabilidad de que aparezcan en los documentos que pertenecen a la categoría en cuestión. Naturalmente, dichas características son los términos que conforman los documentos, y tanto su probabilidad de aparición en general, como la probabilidad de que aparezcan en los documentos de una determinada categoría, pueden obtenerse a partir de los documentos de entrenamiento; para ello se utilizan las frecuencias de aparición en la colección de entrenamiento.

Cuando las colecciones de aprendizaje son pequeñas, pueden producirse errores al estimar dichas probabilidades. Por ejemplo, cuando un determinado término no aparece nunca en esa colección de aprendizaje pero aparece en los documentos a categorizar. Esto implica la necesidad de aplicar técnicas de suavizado, a fin de evitar distorsiones en la obtención de las probabilidades [6].

Con dichas probabilidades, obtenidas de la colección de entrenamiento, podemos estimar la probabilidad de que un nuevo documento, dado que contiene un conjunto determinado de términos, pertenezca a cada una de las categorías. La más probable, obviamente, es a la que será asignado.

El clasificador Bayes Ingenuo combina el modelo de características independientes con una regla de decisión. El clasificador Bayer (la función Classify) se define como:

$$\text{classify}(f_1, \ldots, f_n) = \underset{c}{\operatorname{argmax}}\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c).$$

Figura 4 .- Función Classify

3.3.2.- Algoritmo del Vecino más Próximo y Variantes

El algoritmo del vecino más próximo (Nearest Neighbour, NN) es uno de los más sencillos de implementar. La idea básica es como sigue: si se calcula la similitud entre el documento a clasificar y cada uno de los documentos de entrenamiento, aquél de éstos más parecido estará indicando a qué clase o categoría se debe asignar el documento que se desea clasificar.

Una de las variantes más conocidas de este algoritmo es la del k-nearest neighbour o KNN que consiste en tomar los k documentos más parecidos, en lugar de sólo el primero. Como en esos k documentos los habrá, presumiblemente, de varias categorías, se suman los coeficientes de los de cada una de ellas. La que más puntos acumule, será la candidata idónea. El KNN une a su sencillez una eficacia notable. Obsérvese que el proceso de entrenamiento no es más que la indexación o descripción automática de los documentos, y que tanto dicho entrenamiento como la propia categorización pueden llevarse a cabo con instrumentos bien conocidos y disponibles para cualquiera. De otra parte, numerosas pruebas experimentales han mostrado su eficacia. KNN parece especialmente eficaz cuando el número de categorías posibles es alto, y cuando los documentos son heterogéneos y difusos. Es un método de clasificación no paramétrico, ya que no se hace ninguna suposición distribucional acerca de las variables predictoras. Para inferir la categoría de un ejemplo desconocido, el algoritmo compara ese ejemplo con todos los ejemplos de entrenamiento, calculando la distancia entre ellos. A continuación, la clase mayoritaria de entre los k ejemplos más similares al de entrada es la categoría inferida para el mismo. Generalmente se usa la distancia Euclidiana:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{|C|}(x_{ik} - x_{jk})^2}$$

Figura 5.- Ecuación Euclidiana

En la Figura 6 se muestra un ejemplo del algoritmo KNN para un sistema de dos atributos. En este ejemplo se ve cómo en el proceso de aprendizaje se almacenan todos los ejemplos de entrenamiento. Se han representado los ejemplos de acuerdo a los valores de sus dos atributos y la clase a la que pertenecen (las clases son positivo y negativo). La clasificación consiste en la búsqueda de los k ejemplos (en este caso tres) más cercanos al ejemplo a clasificar. Concretamente, el ejemplo A se clasificaría como negativo, y el ejemplo B como positivo.

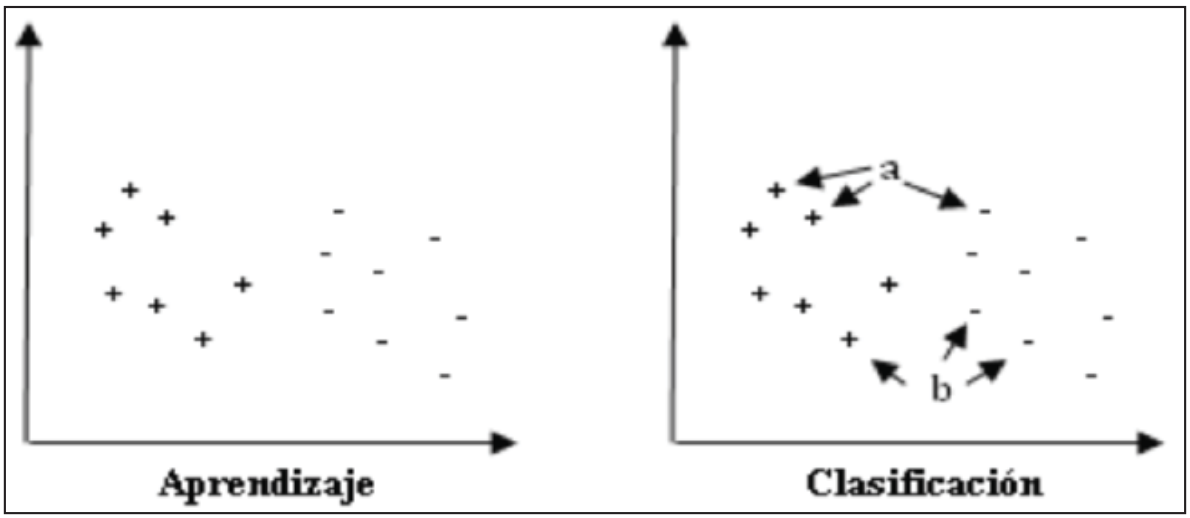


Figura 6.- Algoritmo KNN

3.3.3.- Árboles de Decisión o Clasificación

Es uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado [7]. Como forma de representación del conocimiento, los árboles de clasificación destacan por su sencillez. A pesar de que carecen de la expresividad de las redes semánticas o de la lógica de primer orden, su dominio de aplicación no está restringido a un ámbito concreto sino que pueden utilizarse en diversas áreas: diagnóstico médico, juegos, predicción meteorológica, control de calidad, etc.

Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto de prototipos (documentos). Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada nodo hoja se refiere a una decisión (clasificación).

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar. La metodología a seguir puede resumirse en dos pasos y se esquematiza en las siguientes imágenes (Figuras 7 y 8):

Aprendizaje: Consiste en la construcción del árbol a partir de un conjunto de prototipos. Constituye la fase más compleja y la que determina el resultado final. A esta fase se dedica la mayor parte de la atención.

Clasificación: Consiste en el etiquetado de un patrón, X, independiente del conjunto de aprendizaje. Se trata de responder a las preguntas asociadas a los nodos interiores utilizando los valores de los atributos del patrón X. Este proceso se repite desde el nodo raíz hasta alcanzar una hoja, siguiendo el camino impuesto por el resultado de cada evaluación.

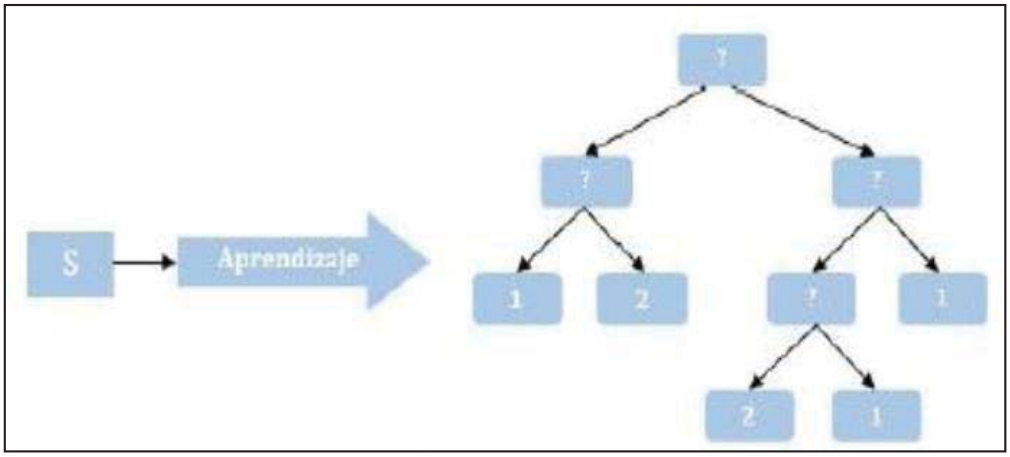


Figura 7.- Árbol de Decisión Aprendizaje

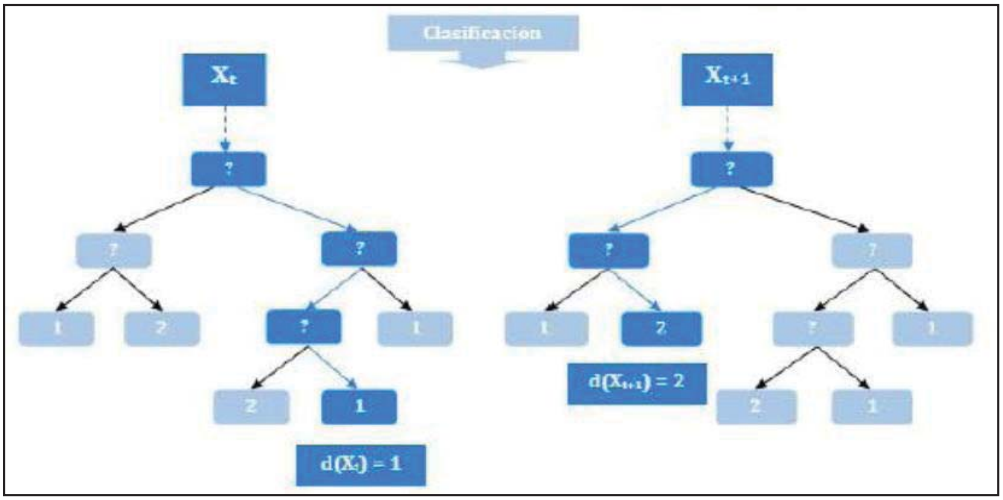


Figura 8.- Árbol de Decisión Clasificación

3.3.4.- Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial o máquinas de vectores de soporte son un conjunto de algoritmos de aprendizaje supervisado. Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento, podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra.

Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase. Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta.

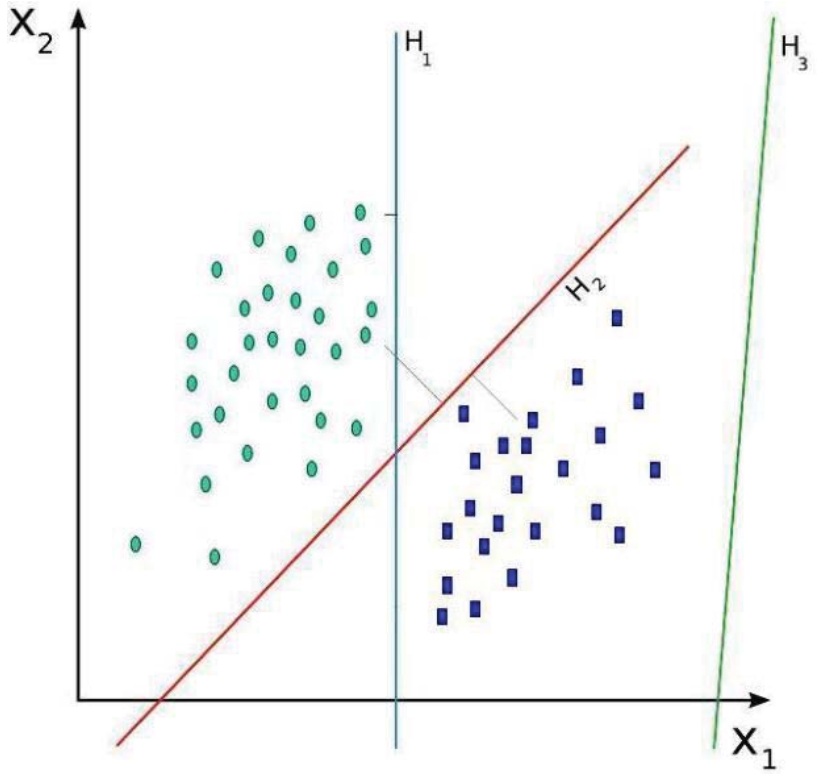


Figura 9.- Máquina de Soporte Vectorial

4.- Twitter y sus Características

4.1.- Antecedentes

Antes de hablar de cómo se puede extraer la polaridad de sentimientos de comentarios emitidos en Twitter, resulta relevante comprender su estructura y características principales.

Después de la aparición de los blogs y respondiendo a la necesidad que tenían los usuarios de poder enviar mensajes de forma rápida y fácil, nace el Microblogging con su principal actor Twitter. En esta nueva forma de comunicación, los usuarios pueden describir un estado o Tweet en un mensaje corto vía Internet, celulares o emails. Esta red social fue lanzada en Octubre del año 2006 y para Abril del 2007 siguiente ya contaba con 94.000 usuarios. Teniendo un crecimiento mantenido para Diciembre del año 2013 Twitter ya contaba con 200 millones de usuarios activos a nivel mundial y 400 millones de comentarios al día [8].

En Twitter los usuarios interactúan y se comunican entre ellos mediante tweets. Estos corresponden a mensajes cortos con un máximo de estrictamente 140 caracteres de largo. Muchos de ellos referencian a otros usuarios de la red social, otro Tweet o a una página Web siguiendo las siguientes convenciones:

4.2.- Estructura de Twitter

Menciones: Si nos queremos referir a algún usuario de Twitter, lo haremos por su código @ xxxxx. Así esta persona, aunque no nos siga, verá en su Twitter que hemos hablado. Es una manera que esta persona nos lea y que, si le interesa lo que decimos nos siga.

Retweet (RT): Un Retweet es un Tweet reenviado, tal como ha llegado o con comentarios adicionales. Se reconoce normalmente por el código RT, aunque no es estrictamente obligatorio. Los programas de gestión de Twitter, ya sea el oficial o los muchos que hay disponibles, identifican los ReTweets.

DM: Es un mensaje directo (DM) a un usuario que sólo verá este usuario. Es completamente confidencial y sólo podemos enviar un DM a una persona que nos siga.

#: Una etiqueta representa un tema, indica un mismo tema sobre el cual cualquier usuario puede opinar simplemente escribiendo un tweet con esta etiqueta al mensaje: por ejemplo #Tema

5.- Arquitectura del Sistema

5.1.- Set de Datos

El conjunto de datos utilizado para el entrenamiento de las técnicas mencionadas anteriormente se encuentra compuesto de 10.000 mil Tweets los cuales fueron entregados por la Empresa Analitic S.A.

Los Tweets se encuentran en su totalidad en idioma español y pertenecen a una cuenta de emergencias de una compañía de electricidad tomados entre Enero y Abril del presente año.

Como se ha mencionado las técnicas indicadas requieren de un conjunto de datos clasificados de forma manual para poder clasificar de una forma más adecuada un set de datos que se encuentre sin catalogar. Es por este motivo que se ha creado un sistema que permita a distintas personas especificar a qué categoría pertenece un Tweet, para así obtener un set de datos inicial lo más objetivo posible.

5.2.- Definición de Categorías

Para el proceso de clasificación de los datos se han distinguidos cinco grandes categorías en las cuales se pueden agrupar los Tweets pertenecientes al Área de Atención de Clientes de una compañía. Estas categorías corresponden a:

Preguntas: En esta categoría se agruparan todos los Tweets que contengan algún enunciado interrogativo que tengan como finalidad conocer algo u obtener alguna información por parte de la empresa.

Sugerencias: En esta categoría se congregaran los Tweets que se proponga alguna idea para mejorar el servicio o algún proceso de la empresa.

Información Usuario: En esta categoría se agruparan aquellos Tweets que estén entregando alguna información relacionada con algún evento o proceso de la empresa.

Otras: En esta categoría se congregaran los Tweets que no se tenga claridad a que categoría anterior pertenecen.

Información Empresa: En esta categoría se agruparan aquellos Tweets que hayan sido publicados por la empresa y algún usuario los haya retwitteado.

5.3.- Proceso de Clasificación

La arquitectura a desarrollar para obtener un clasificador adecuado estará basada en un sistema en cascada que, tras pasar por una fase de preprocesado de las opiniones y entrenamiento del sistema, debe proporcionar la capacidad de clasificar automáticamente nuevos Tweets. Este enfoque, por tanto, se centra en tener un conjunto de Tweets de entrenamiento previamente clasificados, que se usarán para aprender a clasificar nuevos documentos. Para ello, se deben transformar los Tweets de su formato inicial a una representación que pueda ser usada por un algoritmo de aprendizaje para la clasificación.

No todos los elementos que aparecen en los Tweets serán útiles para su clasificación, es decir, hay elementos que por sí mismos no dicen nada del contenido del Tweets en el que se encuentran y que por lo tanto pueden ser eliminados; entre ellos se incluyen por ejemplo los signos de puntuación; también aparecen palabras de uso muy frecuente, palabras que aparecen en una gran cantidad de documentos, lo que hace que su poder discriminatorio sea muy bajo; a este tipo de palabras se les conoce como palabras vacías (StopWords). Una vez que se ha obtenido una representación adecuada de todos los Tweets de entrenamiento (por ejemplo de acuerdo a un modelo de espacio vectorial donde se asigna un peso a las palabras que los conforman) se puede aplicar un método de clasificación (de entre los muchos posibles) para clasificar nuevos elementos.

En la siguiente Figura se puede observar de forma general el proceso de clasificación de un conjunto de documentos.

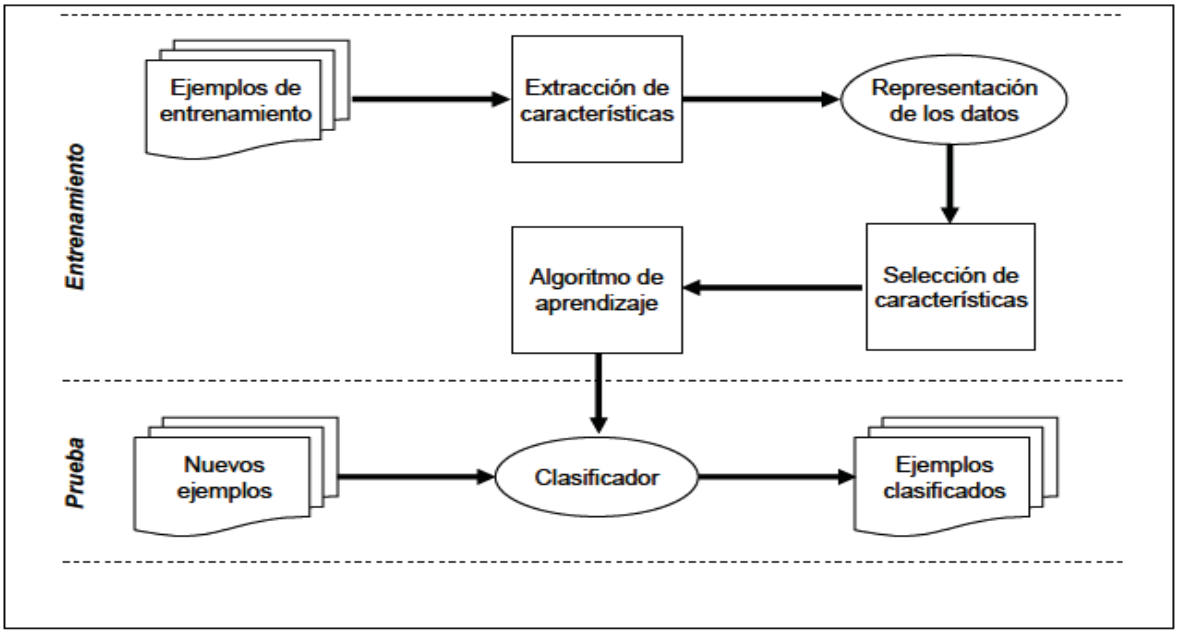


Figura 10.- Fases de un Clasificador Supervisado

5.3.1.- Clasificación de Documentos

El grado de precisión y exactitud que un clasificador automático sea capaz de brindar al clasificar un conjunto de datos dependerá enormemente de que también haya estado entrenado. Por lo que en un principio es sumamente importante tener un conjunto de datos bien clasificados según las categorías definidas anteriormente. Es por este motivo que se ha implementado un pequeño Sitio Web, en el cual un grupo de cinco personas clasifico el conjunto de datos.

Se ha considerado a cinco personas para que la clasificación sea los más objetiva posible. A demás de permitir clasificar únicamente un documento en una sola de las categorías mencionadas, utilizando para ello un sistema de mayoría de votos, vale decir si un Tweet es clasificado por 3 personas en una categoría el Tweet pertenecerá a esa categoría, sin embargo si hay un empate en la clasificación del Tweet este será asignado a la categoría de Otros previa revisión.

A continuación se muestra una tabla de la votación realizada por los cinco usuarios.

Tabla 1.- Votación de Usuarios

	usuario 1	usuario 2	usuario 3	usuario 4	usuario 5
Pregunta	2293	2290	2270	2285	2300
Sugerencia	85	96	90	92	76
Información	6000	6100	6054	5929	6016
Otros	1142	1026	1109	1219	1126
Empresa	480	488	477	475	482
Total	10000	10000	10000	10000	10000

El resultado de la clasificación manual sobre los documentos ha arrojado los siguientes resultados.

Resumen 10000 Tweets Clasificados

Categoria	Total Tweets	Porcentaje %
Pregunta	2279	22.79
Sugerencia	86	0.86
Informacion Usuario	5992	59.92
Otros	1173	11.73
Informacion Empresa	470	4.7

Figura 11 .- Resumen Clasificación Manual de Tweets

Como se puede apreciar en la figura anterior hay una gran cantidad de Tweets que pertenece a la categoría de Información Usuario esto puede presentar un problema al momento de la fase de entrenamiento ya que puede provocar un sesgo por parte de los clasificadores a esta categoría.

5.3.2.- Representación de los Documentos

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.

Este modelo consiste en representar la colección de documentos como una matriz de palabras o términos por documentos [9]. Es decir, cada texto o documento *Dj* es representado por medio de un vector *dj=(W1f…W|t|j)* de términos *W*, donde *T* es el conjunto de palabras del vocabulario de la colección y *Wij* representa un valor numérico que expresa en qué grado el documento *dj* posee el término *Ti*. Frecuentemente, el conjunto *T* es el resultado de filtrar las palabras del vocabulario con respecto a una lista de palabras vacías, éstas son palabras frecuentes que no contienen información semántica (de ahí el nombre de palabras vacías).

5.3.3.- Reducción de Dimensiones

Para poder realizar la clasificación no sólo es necesario obtener el vector de términos mencionado anteriormente, sino que a su vez es fundamental que dicho vector contenga aquellos términos que son más representativos para poder realizar la clasificación con la máxima efectividad posible. Este paso es fundamental para determinar la calidad del clasificador. Hay que seleccionar aquellas palabras clave que aportan significado y descartar aquellas otras que no contribuyen a realizar la distinción entre documentos. Por tanto, es muy conveniente llevar a cabo un proceso de reducción de dimensión (DR, Dimensionality Reduction), obteniendo así un conjunto de términos reducidos. Gracias a esto, se consigue evitar el sobre entrenamiento (overfitting) en el aprendizaje y se aumenta la eficiencia y efectividad del clasificador.

La elección de las palabras clave se realiza descartando aquéllas que aparecen de forma ocasional en el corpus y las que aparecen muy frecuentemente. El motivo es porque si una palabra aparece muy pocas veces en todo el corpus, seguramente se tratará de un error gramatical u otro caso especial que disminuya la calidad del clasificador. Por otra parte, si una palabra aparece de forma habitual en todas las categorías (preposiciones, verbos auxiliares), será demasiado general para ser utilizada para llevar a cabo la clasificación. Los símbolos de puntuación, los números y los caracteres especiales también son descartados, al igual que las palabras que sólo aparecen en un documento o en un número menor que un umbral exceptuando caracteres como (‘?’,’@’).

5.3.3.1.- StopWords o Palabras Vacías

Palabras vacías o StopWords es el nombre que reciben las palabras sin significado como los artículos, pronombres, preposiciones, etc. que han sido filtradas antes del procesamiento de datos.

Se ha generado un diccionario con distintas palabras que han sido categorizadas como vacías ya que no aportan valor al momento de clasificar algún Tweet en alguna de las categorías, También se han incluido aquellas palabras que están presentes en varias categorías.

Algunos ejemplos de estas palabras son:

Tabla 2.- Ejemplos de StopWords

adelante	dado	gueno
ajena	de	haya
alli	ejemplo	otra
ambos	eramos	principalmente
aquel	ex	sta

5.3.3.2.- Lematización

La Lematización es un proceso lingüístico que consiste en, dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. Es decir, el lema de una palabra es la palabra que nos encontraríamos como entrada en un diccionario tradicional: singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos.

Por ejemplo, decir es el lema de dije, pero también de diré o dijéramos; guapo es el lema de guapas; mesa es el lema de mesas etc.

5.3.3.3.- Reducción por Elementos de un Tweet

Otra forma usada para reducir el número de características es en base a analizar el cuerpo de cada mensaje.

Se cambiara toda referencia a alguna URL por lo siguiente “URL_DATA” ya que esta como texto no aporta valor, cabe destacar que al realizar la clasificación de forma manual no fueron visualizados los contenidos de estas, por lo que no infirió en la clasificación del Tweet.

Se cambiara también cada referencia a un tema de interés representado por el signo # por la siguiente palabra tag_data.

Cada Tweet puede tener distintas menciones a usuarios con lo cual tendríamos muchos nombres que puede que no representen alguna importancia significativa en la clasificación de ese Tweet. Por lo cual se ha decido solamente dejar el carácter de mención (@), el cual si puede aportar a la clasificación de ese Tweet. Sin embargo, se han detectado un grupo de usuarios que si poseen importancia al momento de clasificar, estos usuarios corresponden a las cuentas de las compañías. Por lo que se ha decido en esos casos dejar la mención completa, algunos de estos usuarios son: @cged_sos, @conafe, @conafe_sos y @elecda_sos entre otros.

5.3.3.4.- Reducción por Contexto de los Tweets

Analizando el contexto de los Tweets los cuales corresponden a mensajes entre los clientes de una compañía eléctrica y viceversa. Se ha analizado que en muchas ocasiones para realizar de mejor manera la atención, los usuarios indican sus nombres los cuales no aportan por si un grado de importancia que permitan predecir alguna categoría, sin embargo el hecho de dar sus datos personales puede aportar a predecir la categoría a la que pertenece el Tweet. Por lo que se ha creado la siguiente regla: Se ha creado un diccionario el cual está formado por nombres y apellidos, cada vez que se encuentre un nombre o apellido que este registrado en el diccionario se remplazara ese nombre o apellido por la palabra nombre_cliente de esta manera se logra lo explicado anteriormente . Así mismo ocurre con los datos numéricos ya sean RUT, Teléfono o el número de cliente (Identificación del usuario en el sistema de la compañía), los cuales han sido reemplazos por la siguiente palabra: Info_Numero.

5.3.3.5.- Aplicando la Reducción de Dimensiones

Para reducir las dimensiones se han aplicado todas las formas a cada Tweet antes de ser procesado lo cual ha dado los siguientes resultados.

rt @annytagr: @elecda_sos corte de luz en el sector norte, aque hora habrá una reposición? @haroldrivasm @chrischile #antofaga…
Tweet Normal

rt @ @elecda_sos corte luz sector norte aque hora habra reposicion ? @ @ tag_data
Tweet Limpio

avisar @cged_sos #talca https://t.co/jvxug22k44
Tweet Normal

avisar @cged_sos tag_data url_data
Tweet Limpio

hola baja voltaje en sector feria las pulgas. lorenzo arenas 7465
Tweet Normal

baja voltaje sector feria pulgas dato_cliente dato_cliente
Tweet Limpio

5.3.4.- Asignación de Pesos

Según el modelo de espacio vectorial, un documento puede ser considerado como un vector Dj = {W1j,…..W|T|j}, donde Wkj es un valor numérico que expresa la importancia de la palabra k en el documento j. A continuación se detallaran tres métodos que permiten determinar el peso de las palabras que conforman un determinado documento.

5.3.4.1.- Representación Binaria

En cada posición del vector se indica la presencia (1) o no presencia (0) de una palabra correspondiente a esa posición.

Documento	w_1	w_2	w_3
d_1	1	0	1
d_2	1	0	0
d_3	0	1	0
d_4	0	1	1
d_5	1	1	1
d_6	0	0	0
d_7	1	1	0

Figura 12.- Ejemplo Representación Binaria

5.3.4.2.- Representación TF-IDF

La ponderación de los términos es el proceso que tiene como finalidad conocer la importancia de los términos para representar un documento y permitir su posterior recuperación. Esto implica que se debe determinar el poder de resolución de los términos de la colección, o lo que es lo mismo, la capacidad de los términos para representar el contenido de los documentos en la colección, que permitan identificar cuáles son relevantes o no. Al valor e índice que es capaz de determinar este extremo se le denomina "peso del término" o "ponderación del término" y su cálculo implica determinar la "Frecuencia de aparición del término TF" y la "Frecuencia inversa del documento para un término IDF".

El factor TF es la suma de todas las ocurrencias o el número de veces que aparece un término en un documento. A este tipo de frecuencia de aparición también se la denomina "Frecuencia de aparición relativa" por que atañe a un documento en concreto y no a toda la colección. El valor de este factor tiene tres casos elementales:

- Frecuencia de aparición TF baja, Representa una representividad elevada.
- Frecuencia de TF media, Representa un término relativamente común.
- Frecuencia de aparición TF alta, Representa muy baja representatividad.

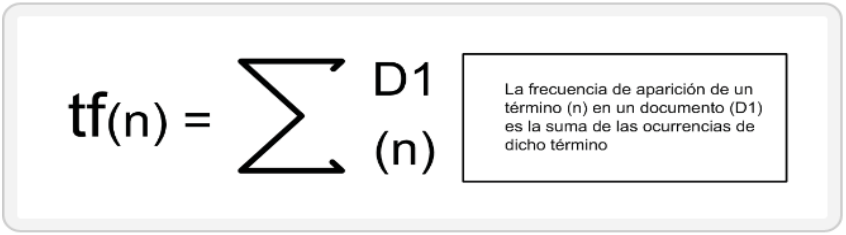


Figura 13.- Fórmula para calcular TF

El factor IDF de un término es inversamente proporcional al número de documentos en los que aparece dicho término. Esto significa que cuanto menor sea la cantidad de documentos, así como la frecuencia absoluta de aparición del término, mayor será su factor IDF y a la inversa, cuanto mayor sea la frecuencia absoluta relativa a una alta presencia en todos los documentos de la colección, menor será su factor discriminatorio.

El valor de este factor tiene tres casos elementales:

- Poder discriminatorio bajo, Término genérico aparece en varios documentos.
- Valor IDF medio, Representa un término relativamente común.
- Poder discriminatorio alto, Término aparece en pocos documentos.

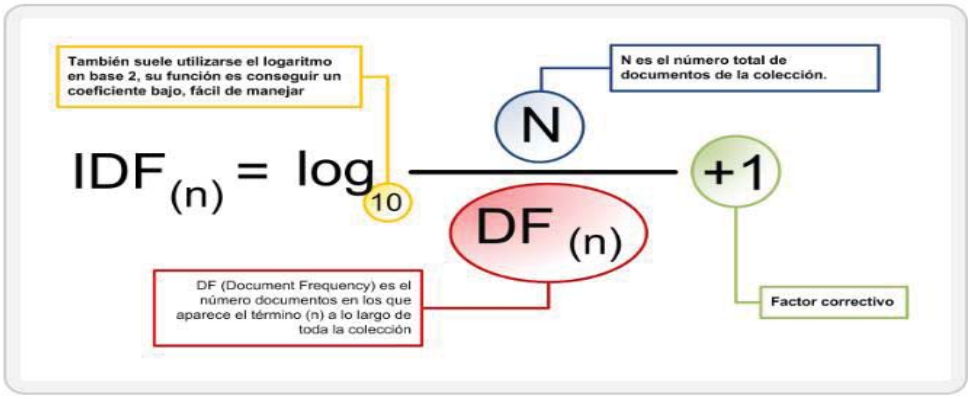


Figura 14.- Fórmula para Calcular IDF

La Ponderación TF-IDF, es el peso de un término en un documento es el producto de su frecuencia de aparición en dicho documento (TF) y su frecuencia inversa de documento (IDF) tal como refleja la Figura 19.

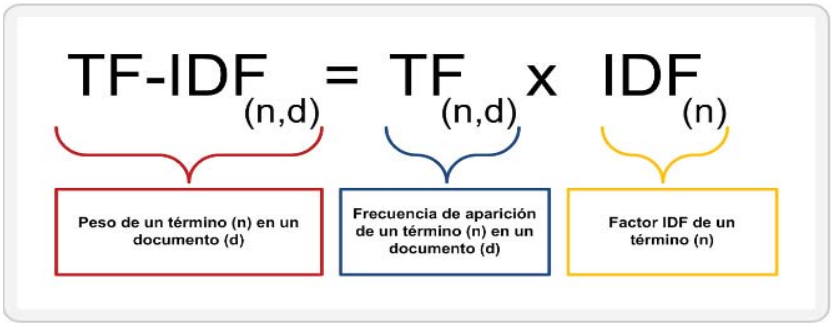


Figura 15.- Fórmula para Calcular TF-IDF

5.3.4.3.- Representación TF-RFL

Corresponde a la relevancia de la frecuencia de una categoría (etiqueta), el cual es una representación propuesta por [10], la que constituye una nueva representación para el problema de múltiples categorías.

$$tf - rfl_{tdl} = f_{td} \log_2\left(2 + \frac{a_{t,l}}{\max(1, mean(a_{t,\lambda_{j/l}}))}\right)$$

Figura 16.- Formula TF-RFL

En donde $mean(at, \lambda j/l)$ es el número promedio de documentos que contienen el término *t* para cada documento clasificado en categorías diferentes a *l*.

5.3.5.- Entrenamiento y Clasificación

Para desarrollar el clasificador se debe seleccionar un número apropiado de documentos previamente clasificados a los que se denomina conjunto de entrenamiento. Este conjunto, una vez transformado adecuadamente de su formato inicial a una representación que pueda ser tratada por el algoritmo de aprendizaje, servirá para entrenar el sistema y capacitarle para decidir la clase a la que pertenecerán nuevos documentos entrantes cuya clase se desconoce.

5.3.5.1.- Validación Cruzada

Para realizar este proceso se ha utilizado un método que se llama Validación Cruzada**,** Este método se inicia mediante el fraccionamiento de un conjunto de datos en un número K de particiones (generalmente entre 5 y 10) llamadas pliegues. La validación cruzada luego itera entre los datos de evaluación y entrenamiento K veces, de un modo particular. En cada iteración de la validación cruzada, un pliegue diferente se elige como los datos de evaluación. En esta iteración, los otros pliegues K−1 se combinan para formar los datos de entrenamiento. Por lo tanto, en cada iteración tenemos (K−1)/K de los datos utilizados para el entrenamiento y 1/K utilizado para la evaluación. Cada iteración produce un modelo, y por lo tanto una estimación del rendimiento de la generalización, por ejemplo, una estimación de la precisión. Una vez finalizada la validación cruzada, todos los ejemplos se han utilizado sólo una vez para evaluar pero K−1 veces para entrenar. En este punto tenemos estimaciones de rendimiento de todos los pliegues y podemos calcular la media y la desviación estándar de la precisión del modelo. Veamos un ejemplo

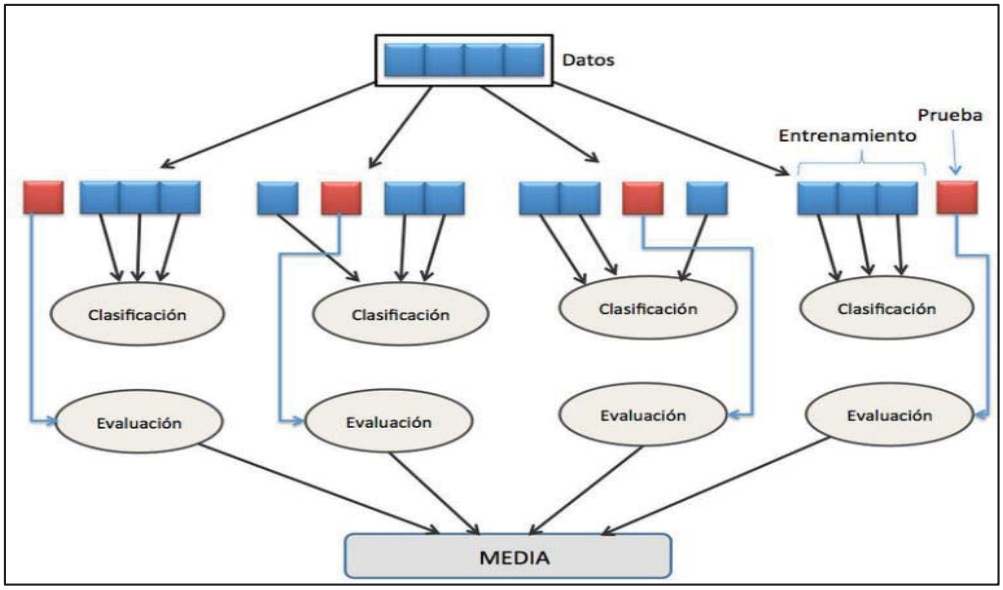


Figura 17.- Ejemplo Validación Cruzada

En este proyecto, se utilizara el iterador StratifiedKFold que nos proporciona Scikit-learn. Este iterador es una versión mejorada de la validación cruzada, ya que cada pliegue va a estar estratificado para mantener las proporciones entre las clases del conjunto de datos original, lo que suele dar mejores estimaciones del sesgo y la varianza del modelo. Se ha utilizado un K = 5 con 10 iteraciones en las cuales se han evaluado los métodos que se detallaran a continuación.

5.3.5.2.- Arboles de Decisión

Los árboles de decisión son un método de clasificación no paramétrico supervisado utilizado para el aprendizaje de clasificación y regresión . El objetivo es crear un modelo que predice el valor de una variable objetivo, aprendiendo las reglas de decisión simples inferidas a partir de las características de los datos.

El algoritmo utilizado es una versión optimizada del algoritmo CART, el cual consiste en construir los arboles binarios usando la función y el umbral que dio la mayor ganancia de información en cada nodo.

5.3.5.3.- K-Nearest Neighbors

El principio detrás de los métodos de vecinos más cercanos es encontrar un número predefinido de muestras de entrenamiento más cercanos en la distancia hasta el nuevo punto, y predecir la etiqueta de estos. El número de muestras puede ser una constante (k-más cercano aprendizaje vecino) definida por el usuario, o variar en función de la densidad local de los puntos (a base de radio de aprendizaje vecino). La distancia puede, en general, ser cualquier medida métrica: distancia euclidiana estándar es la opción más común. El valor de K usado en las pruebas de este algoritmo se ha fijado en 5. K=5.

5.3.5.4.- Bernoulli Naive Bayes

Este método es una variante del método Naive Bayes Ingenuo explicado anteriormente en la cual se implementa el algoritmo bayesiano para los datos que se distribuye de acuerdo a las distribuciones de Bernoulli multivariados; es decir, puede que haya muchas características, pero cada uno se supone que es una variable binaria de valor (Bernoulli, boolean). La regla de decisión para Bernoulli Bayes ingenuo se basa en:

$$\boxed{P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)}$$

En la cual se penaliza explícitamente la no ocurrencia de una característica i que es un indicador para la clase y , Donde la variante multinomial sería simplemente ignorar una característica no se produzca.

5.3.5.5.- Máquina de Soporte Vectorial

Las Máquinas de vectores de soporte (SVMs) son un conjunto de métodos de aprendizaje supervisado utilizados para la clasificación , regresión y la detección de valores atípicos.

El núcleo utilizado en el proceso de experimentación será un kernel Lineal, el cual implementará la estrategia Uno Contra Todos, la cual consiste en la adaptación de un clasificador por clase. Para cada clasificador, la clase estará compitiendo contra las otras categorías. Este método además de su eficiencia computacional aporta un gran grado de interpretabilidad. Dado que cada clase está representada por un único y un clasificador.

5.3.6.- Medidas de Evaluación

Otro punto importante y más bien final, es la evaluación de los algoritmos, ya que no siempre entregan una efectividad del 100%, por ende se necesitan métricas capaces de dar información relevante al desempeño de la clasificación.

Las métricas con las que se trabaja y se analiza el desempeño del algoritmo son las siguientes:

5.3.6.1.- Matriz de Confusión

En el campo de la inteligencia artificial, una matriz de confusión es una herramienta visual que se utiliza en el aprendizaje supervisado, cada columna que posee representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias de la clase real. Uno de los principales beneficios de las matrices de confusión es que facilitan ver si el sistema se confunde entre clases.

	Clasificado como A	Clasificado como B
Real A	Verdadero Positivo (VP)	Falso Negativo (FN)
Real B	Falso Positivo (FP)	Verdadero Negativo (VN)

Figura 18.- Matriz de Confusión

5.3.6.2.- Accuracy

Hace referencia a la conformidad de un valor medido con su valor verdadero, es decir, se refiere a cuán cerca del valor real se encuentra el valor medido. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación.

$$Accuracy = \frac{VP + VN}{VP + FP + FN + VN}$$

5.3.6.3.- F1-Score

Es la medida de precisión que tiene un test. Se emplea en la determinación de un valor único ponderado de la precisión y la exhaustividad (Recall).

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5.3.6.4.- Precisión

La precisión es el ratio entre el número de documentos relevantes recuperados entre el número de documentos recuperados. De esta forma, cuanto más se acerque el valor de la precisión al valor nulo, es decir nulo el valor del denominador, mayor será el número de documentos recuperados que no consideren relevantes. Si por el contrario, el valor de la precisión es igual a uno, se entenderá que todos los documentos recuperados son relevantes.

$$Precision = \frac{VP}{VP + FP}$$

5.3.6.5.- Recall

Este ratio viene a expresar la proporción de documentos relevantes recuperados, comparado con el total de los documentos que son relevantes existentes en la base de datos, con total independencia de que éstos, se recuperen o no. Si el resultado de esta fórmula arroja como valor 1, se tendrá la exhaustividad máxima posible, y esto viene a indicar que se ha encontrado todo documento relevante que residía en la base de datos, por lo tanto no se tendrá ni ruido, ni silencio informativo: siendo la recuperación de documentos entendida como perfecta. Por el contrario en el caso que el valor de la exhaustividad sea igual a cero, se tiene que los documentos obtenidos no poseen relevancia alguna.

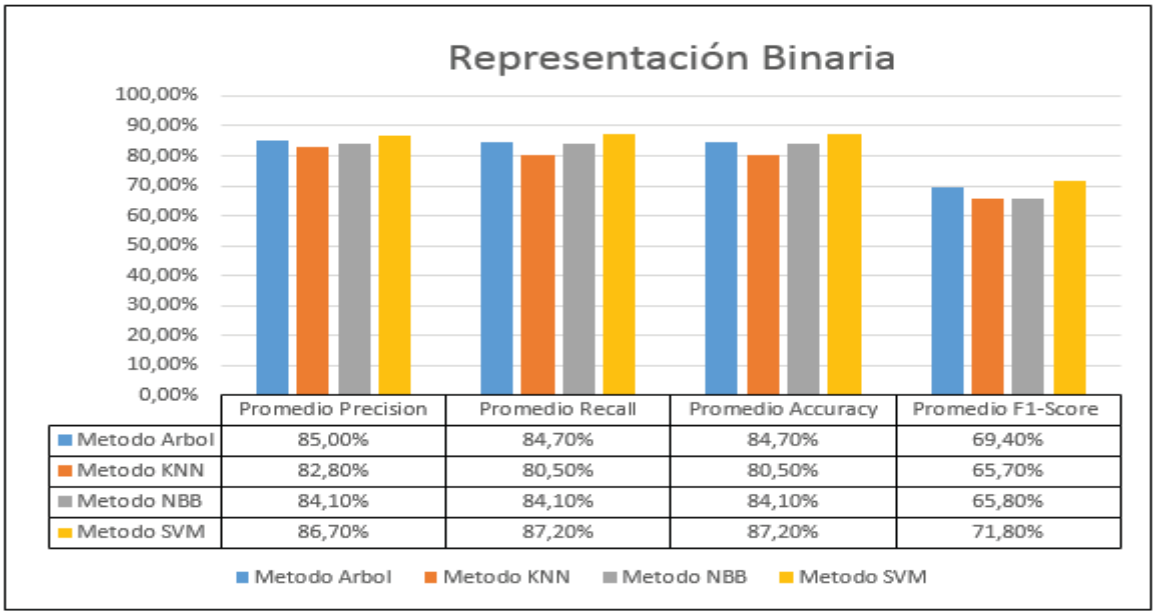
$$Recall = \frac{VP}{VP + FN}$$

6.- Experimentación de Clasificadores

El corpus generado tras realizado todos los procesos del capítulo anterior ha quedado formado por 2845 palabras con las cuales se ha formado el vector representativo de cada Tweet. Una vez generado el proceso de creación del vector se han realizado tres asignaciones de pesos para realizar experimentos de forma individual.

Los experimentos que fueron llevados a cabo fueron los siguientes, como se mencionó en el apartado anterior se ha utilizado el método de Validación Cruzada para realizar las pruebas sobre las distintas representaciones utilizadas, las cuales consisten en las siguientes.

6.1.- Resultados Representación Binaria



	Promedio Precision	Promedio Recall	Promedio Accuracy	Promedio F1-Score
Metodo Arbol	85,00%	84,70%	84,70%	69,40%
Metodo KNN	82,80%	80,50%	80,50%	65,70%
Metodo NBB	84,10%	84,10%	84,10%	65,80%
Metodo SVM	86,70%	87,20%	87,20%	71,80%

Figura 19.- Grafico Representación Binaria

Tabla 3- Desviación Estándar Representación Binaria

Desviación Estandar	Metodo Arbol	Metodo KNN	Metodo NBB	Metodo SVM
Precisión	0,002	0,001	0,001	0,002
Recall	0,002	0,002	0,001	0,002
Accuracy	0,002	0,002	0,001	0,002
F1-Score	0,004	0,005	0,001	0,005

A continuación se mostrara una comparativa de las matrices de confusión generadas en una de las iteraciones del proceso de validación cruzada mediante las técnicas utilizadas en el proceso de clasificación, utilizando la representación Binaria.

METODO ARBOL						
10	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	373	1	65	16	0
	Sugerencia	2	2	7	6	0
	Información	63	7	1017	106	5
	Otros	8	1	42	183	0
	Empresa	3	2	13	1	75

METODO KNN						
10	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	288	3	137	27	0
	Sugerencia	0	0	13	4	0
	Información	21	1	1019	155	2
	Otros	3	1	40	190	0
	Empresa	3	0	18	5	68

METODO NBB						
10	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	380	0	50	25	0
	Sugerencia	6	0	2	9	0
	Información	81	0	1044	72	1
	Otros	11	0	51	172	0
	Empresa	8	0	7	3	76

METODO SVM						
10	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	385	3	55	11	1
	Sugerencia	1	1	7	8	0
	Información	22	2	1102	35	4
	Otros	7	1	64	162	0
	Empresa	3	1	10	0	80

Figura 20.- Comparativa Matriz Confusión Representación Binaria

6.2.- Resultados Representación TF-IDF

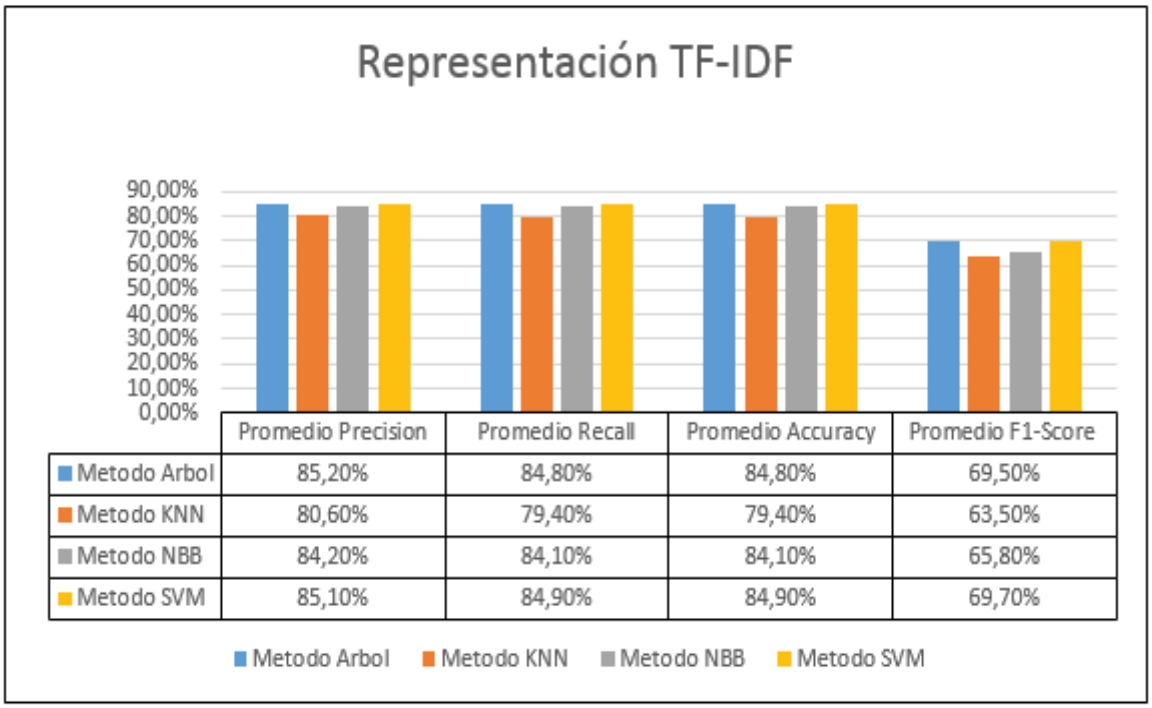


Figura 21.- Grafico Representación TF-IDF

Tabla 4.- Desviación Estándar Representación TF-IDF

Desviación Estandar	Metodo Arbol	Metodo KNN	Metodo NBB	Metodo SVM
Precisión	0,002	0,002	0,002	0,003
Recall	0,001	0,001	0,001	0,002
Accuracy	0,001	0,001	0,001	0,002
F1-Score	0,005	0,002	0,001	0,006

A continuación se mostrara una comparativa de las matrices de confusión generadas en una de las iteraciones del proceso de validación cruzada mediante las técnicas utilizadas en el proceso de clasificación, utilizando la representación TF-IDF.

METODO ARBOL

5	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	404	1	45	5	0
	Sugerencia	0	2	13	2	0
	Información	57	5	1058	71	7
	Otros	5	5	52	172	0
	Empresa	5	0	2	3	84

METODO KNN

5	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	268	0	168	19	0
	Sugerencia	0	2	12	3	0
	Información	24	1	1095	76	2
	Otros	6	0	75	153	0
	Empresa	1	0	15	0	78

METODO NBB

5	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	376	0	64	15	0
	Sugerencia	2	0	12	3	0
	Información	61	0	1062	72	3
	Otros	13	0	65	156	0
	Empresa	3	0	5	0	86

METODO SVM

5	Iteración	Clasificadas				
5	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	394	5	42	10	4
	Sugerencia	5	3	7	2	0
	Información	64	13	1065	47	9
	Otros	12	2	67	151	2
	Empresa	1	0	4	0	89

Figura 22.- Comparativa Matriz Confusión Representación TF-IDF

6.3.- Resultados Representación TF-RFL

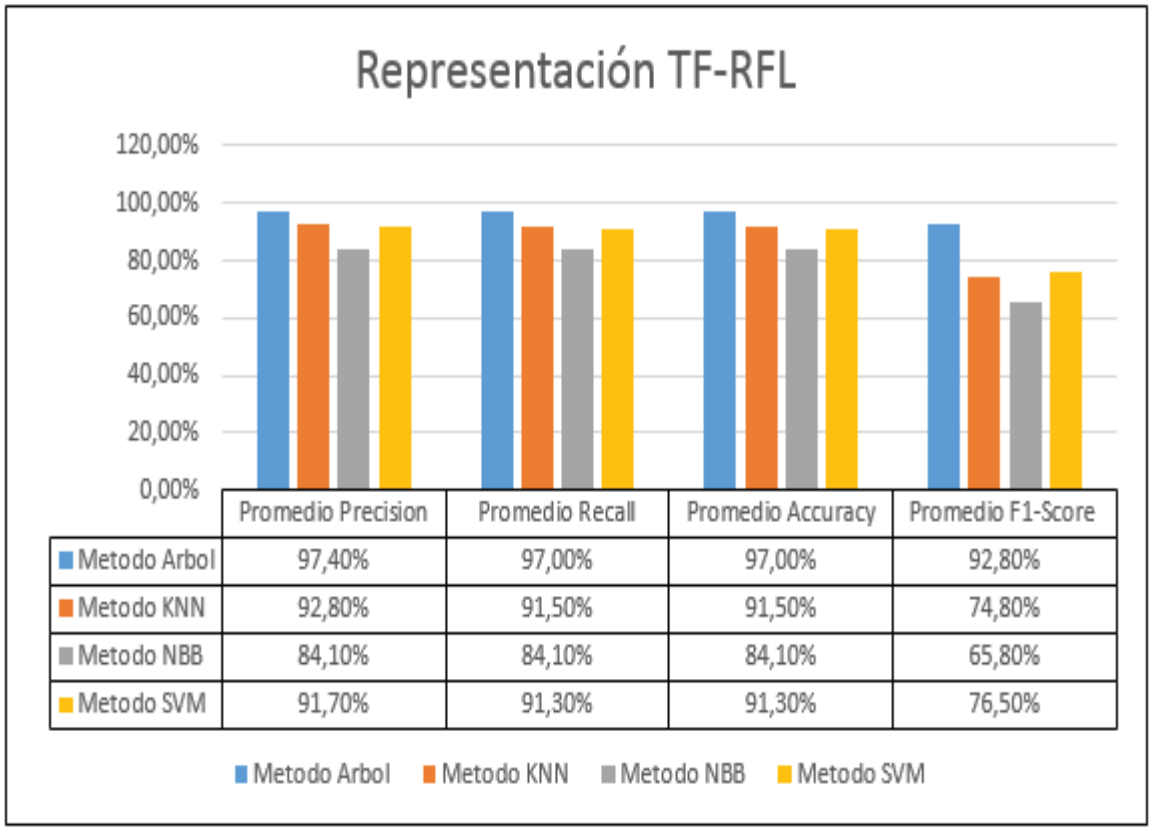


Figura 23.- Grafico Representación TF-RFL

Tabla 5.- Desviación Estándar Representación TF-RFL

Desviación Estandar	Metodo Arbol	Metodo KNN	Metodo NBB	Metodo SVM
Precisión	0,001	0,001	0,001	0,001
Recall	0,001	0,001	0,001	0,001
Accuracy	0,001	0,001	0,001	0,001
F1-Score	0,005	0,004	0,002	0,005

A continuación se mostrara una comparativa de las matrices de confusión generadas en una de las iteraciones del proceso de validación cruzada mediante las técnicas utilizadas en el proceso de clasificación, utilizando la representación TF-RFL.

METODO ARBOL						
1	Iteración	Clasificadas				
1	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	453	0	3	0	0
	Sugerencia	1	13	0	4	0
	Información	4	0	1168	27	0
	Otros	0	0	3	232	0
	Empresa	0	1	0	1	92

METODO KNN						
1	Iteración	Clasificadas				
1	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	422	3	18	13	0
	Sugerencia	0	1	5	12	0
	Información	7	3	1109	77	3
	Otros	5	4	5	221	0
	Empresa	0	1	4	13	76

METODO NBB						
1	Iteración	Clasificadas				
1	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	377	0	64	15	0
	Sugerencia	6	0	8	4	0
	Información	73	0	1035	91	0
	Otros	19	0	32	184	0
	Empresa	4	0	7	2	81

METODO SVM						
1	Iteración	Clasificadas				
1	Fold	Pregunta	Sugerencia	Información	Otros	Empresa
Real	Pregunta	429	2	21	4	0
	Sugerencia	1	2	8	7	0
	Información	23	11	1113	47	5
	Otros	5	2	31	196	1
	Empresa	1	1	5	2	84

Figura 24.- Comparativa Matriz Confusión Representación TF-RFL

6.4.- Comparativa Métodos y Representaciones

A continuación se realizara una comparativa de la precisión de los cuatro métodos utilizados con las distintas representaciones implementadas.

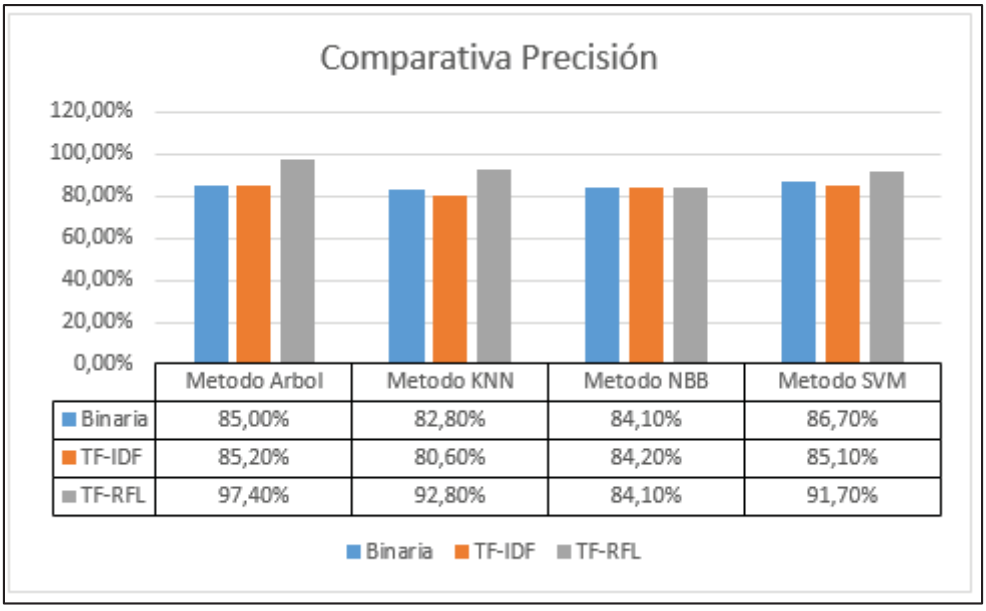


Figura 25.- Grafico Comparativa Resultados Precisión

A continuación se realizara una comparativa de la Accuracy de los cuatro métodos utilizados con las distintas representaciones implementadas.

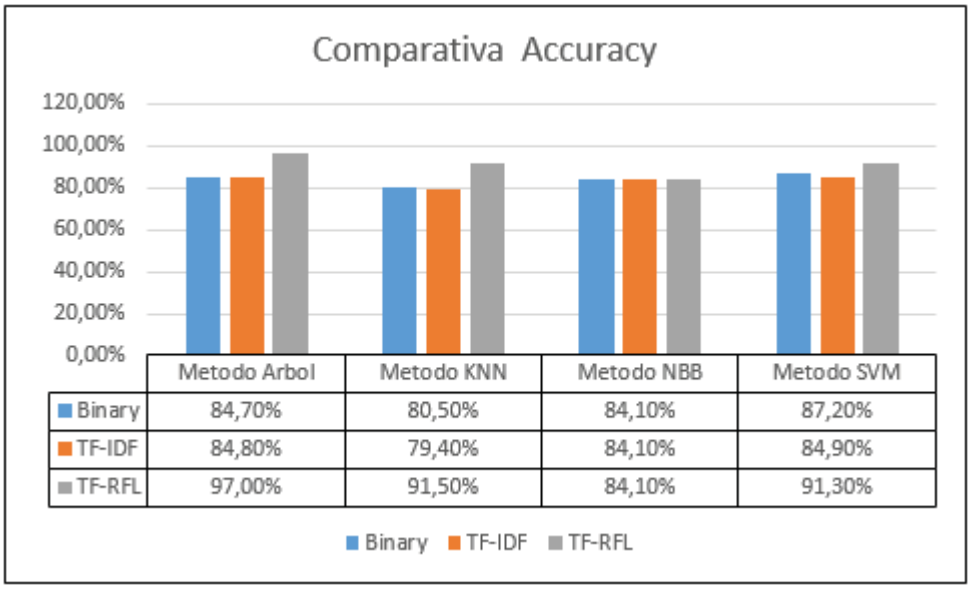


Figura 26.- Grafico Comparativa Resultados Accuracy

A continuación se realizara una comparativa del Recall de los cuatro métodos utilizados con las distintas representaciones implementadas.

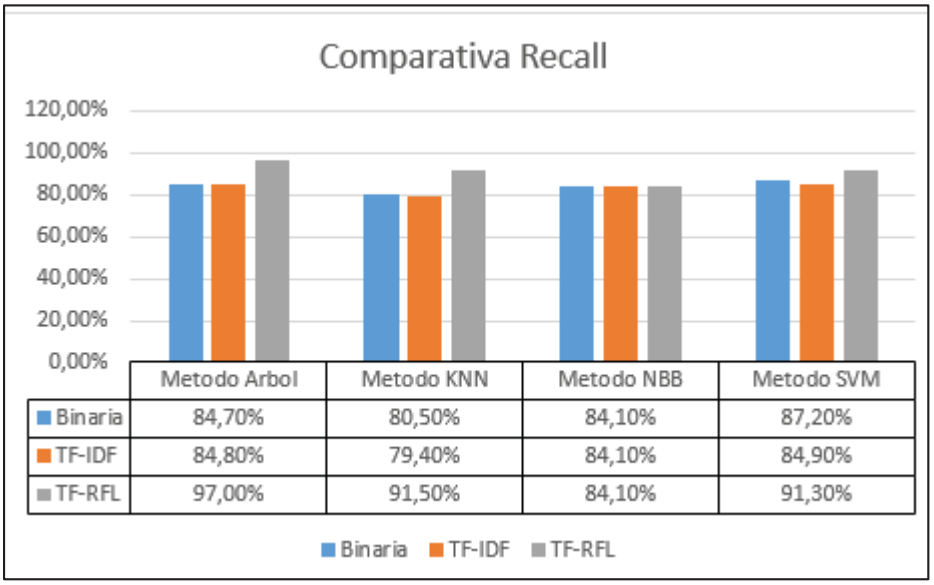
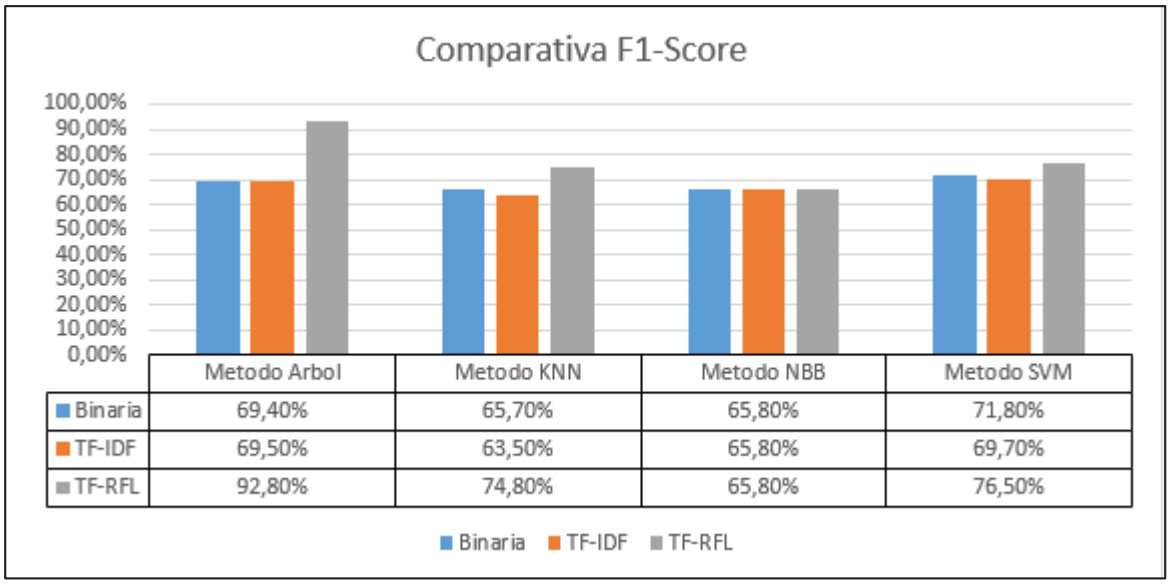


Figura 27.- Grafico Comparativa Resultados Recall

A continuación se realizara una comparativa del F1-Score de los cuatro métodos utilizados con las distintas representaciones implementadas.



	Metodo Arbol	Metodo KNN	Metodo NBB	Metodo SVM
Binaria	69,40%	65,70%	65,80%	71,80%
TF-IDF	69,50%	63,50%	65,80%	69,70%
TF-RFL	92,80%	74,80%	65,80%	76,50%

Figura 28.- Grafico Comparativa Resultados F1-Score

6.5.- Experimentación Semi-Supervisado

Se han realizado las siguientes pruebas para identificar que tan bien funciona el sistema para clasificar nuevos datos que están sin etiquetar, para lograr esto se han sacado una cantidad de 3000 Tweets en total para experimentar. La cantidad de Tweets por categoría se detalla a continuación:

Tabla 6.- Resumen Asignación de Tweets por Categoría

Categoría	Cantidad Tweets
Pregunta	600
Sugerencia	80
Información Usuario	1700
Otros	500
Información Empresa	120

Una vez separado este conjunto de Tweets se ha generado un nuevo corpus con todos aquellos Tweets que no han sido seleccionados, dando como resultado un corpus de 2455, 390 palabras menos que el corpus generado con todos los Tweets. En la creación del corpus se ha introducido un nuevo elemento ([*]), el cual será utilizado para hacer distinción de aquellas palabras que no se encuentren registradas en el corpus y aparezcan en los Tweets que han sido seleccionados para realizar las pruebas mediante la representación TF-IDF.

Una vez generado los vectores representativos se ha elaborado un sistema de votación que funciona de la siguiente manera: Se han utilizado los 7000 Tweets como entrenamiento para las técnicas utilizadas y los 3000 que fueron seleccionados son utilizados como test. El sistema generado consiste en un algoritmo de votación que utiliza las 3 técnicas mencionadas para clasificar cada uno de los Tweets seleccionados. Este sistema de votación funciona de la siguiente manera cada Tweet es clasificado por las 4 técnicas pero únicamente el Tweet será clasificado en alguna categoría si al menos 3 técnicas han predicho la misma categoría para un determinado Tweet.

Una vez realizado el proceso de votación se han obtenido los siguientes resultados:

Tabla 7.- Resumen Tweets Clasificados

Detalle	Cantidad
Total Asignadas	2017
Total Sin Asignar	983

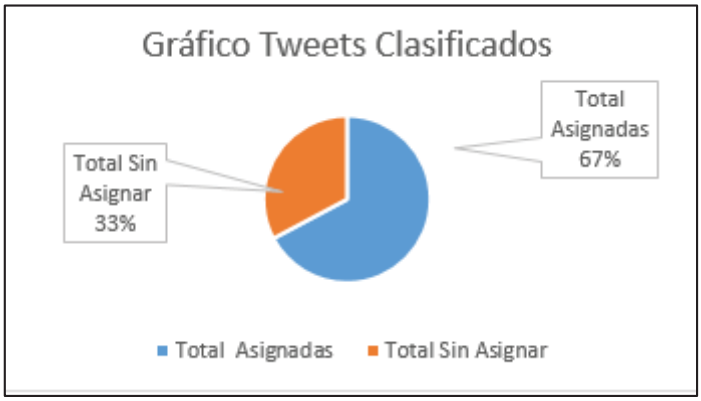


Figura 29.- Grafico Tweets Clasificados

Una vez obtenido la cantidad de datos que el sistema fue capaz de clasificar es importante saber qué porcentaje de esa clasificación fue clasificado satisfactoriamente, a continuación se detallaran esos resultados:

Tabla 8.- Resumen Tweets Clasificados

Detalle	Cantidad
Total Correctas Asignadas	1043
Total Erradas Asignadas	974

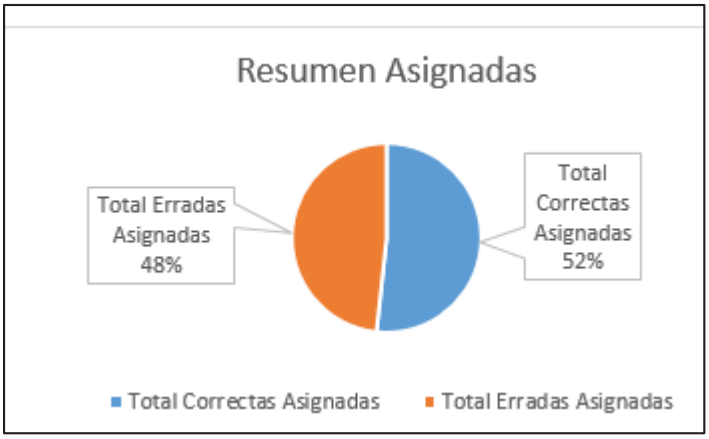


Figura 30.- Resumen Asignadas

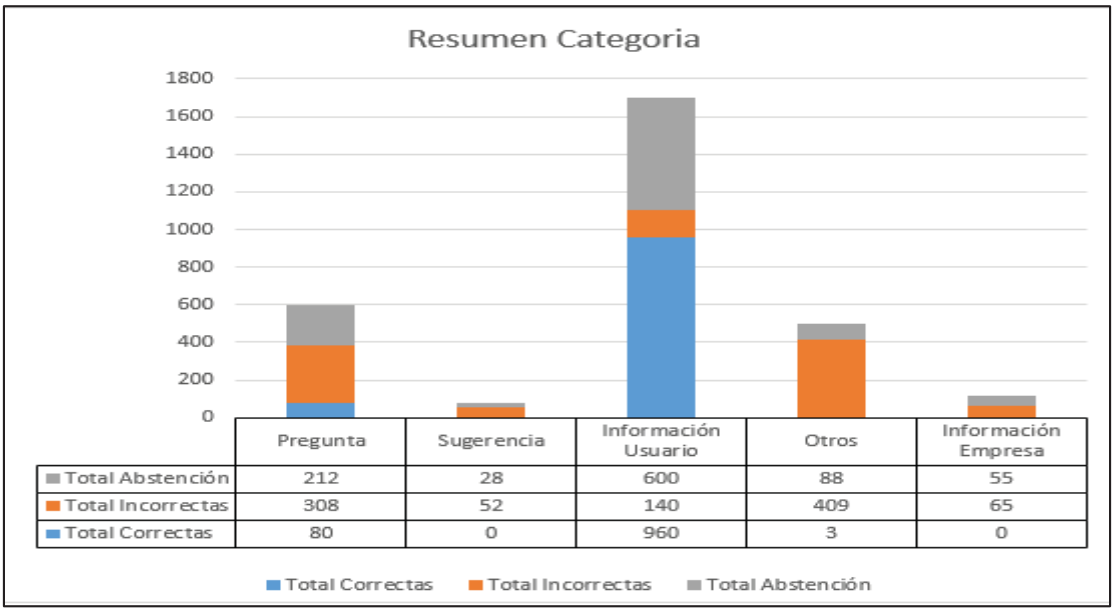


Figura 31.- Grafico Resumen Tweets Clasificados

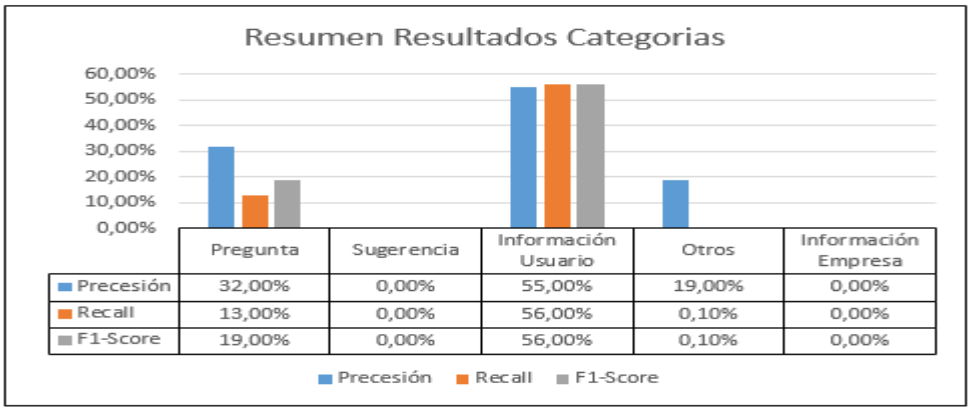


Figura 32.- Resumen Métricas por Categoría

A continuación se mostrara la matriz confusión de las pruebas realizadas donde la categoría abstención son aquellos Tweets que el sistema de votación no fue capaz de clasificar.

Tabla 9.- Matriz Confusión Categorías

	Abstencion	Pregunta	Sugerencia	Usuario	Otros	Empresa
Abstencion	0	0	0	0	0	0
Pregunta	212	80	0	306	2	1
Sugerencia	28	7	0	45	0	0
Usuario	600	129	0	960	11	0
Otros	88	11	0	398	3	0
Empresa	55	23	0	42	0	0

7.- Respuestas Automatizadas

7.1.- Antecedentes

Las respuestas automáticas son mensajes creados de forma automatizada, los cuales buscan responder solicitudes comunes que puedan tener los usuarios respecto a los servicios entregados por una compañía o plataforma. La utilización más común de esta medida es la creación de un apartado en los sistemas digitales denominado Preguntas Frecuentes en el cual se busca dar respuestas a esas inquietudes.

Actualmente los sistemas de mensajería como Twitter, Facebook, WhatsApp entre otros servicios están creando respuestas automáticas a mensajes enviados por los usuarios. La finalidad de estas respuestas automáticas, es especialmente para negocios, autónomos y empresas, el avisar a un cliente o potencial cliente de que no vas a poder contestar en un determinado tiempo puede evitar que se sienta abandonado y eso puede ser la diferencia entre retenerlo o que se te escape a la competencia.

7.2.- Identificación de Respuestas Automáticas

Una vez clasificado los Tweet correspondientes se han analizado si existen consultas generales que los distintos usuarios realizaban a las cuenta de la compañía y que posean respuestas por parte de la empresa, tras el análisis se han identificado tres categorías las cuales poseen respuestas comunes a las diversos mensajes recibidos. Estas categorías corresponden a las siguientes:

Categoría Otros: En esta categoría se ha identificado un grupo de Tweet, en los cuales los usuarios agradecían la gestión de la compañía, que han sido respondidos de una forma similar por parte de la empresa. Por lo que se ha creado una respuesta tipo para responder a ese tipo de mensajes. La respuesta creada tendrá el nombre de Respuesta de Agradecimiento.

Categoría Información Usuarios: En esta categoría se ha identificado un elemento en particular, el cual corresponde a la dirección URL ya sea de una imagen, página web o archivo entre otros. Los cuales se ha verificado que el personal a cargo de revisar los distintos Tweets no visualiza, ya sea por el peligro que pueda llevar abrir una URL desconocida o el tiempo que tiempo que disponen para responder a cada mensaje. Por lo que se ha creado una respuesta para responder a estos tipos de mensaje cuando únicamente este presente una URL. La respuesta creada tendrá el nombre de Respuesta URL.

Categoría Pregunta: En esta categoría se han identificado cinco preguntas las cuales tienen respuestas similares por parte de la compañía. Por lo tanto se han creado cinco respuestas para responder a estas consultas. Las cuales se detallaran a continuación:

- **Respuesta Alumbrado Público:** Esta respuesta responde las consultas realizadas sobre quien es el encargado de realizar las mantenciones o arreglos al alumbrado público.

- **Respuesta Pregunta ¿Qué es DM?:** Esta respuesta responde a las consultas realizadas, en las cuales se pregunta que es un DM**.**

- **Respuesta a Consultas Generales:** Esta respuesta está ligada al contexto que tienen las cuentas analizadas las cuales corresponden como se ha mencionado anteriormente a una cuenta de emergencia de una compañía eléctrica. Por lo cual muchas consultas generales como por ejemplo: ¿cuál es el horario de una sucursal?, ¿cuál es el valor del kw/h? entre otras no están en la finalidad de ser tratadas por las cuentas analizadas.

- **Pregunta Numero Emergencias:** Esta respuesta está relacionada con las consultas en las cuales se pregunta un número de atención de emergencias.

- **Pregunta Operatividad:** Esta respuesta está relacionada con las consultas en las cuales se pregunta si la compañía se encuentra operativa.

7.3.- Reglas para Respuestas

Para realizar el proceso de respuestas automatizadas se ha realizado el siguiente proceso se ha obtenido un corpus por cada respuesta identificada, una vez generado dicho corpus se ha analizado su contenido dando un factor de importancia de dicho termino.

Una vez asignado cada factor a cada elemento del corpus se generara una regla básica que permitirá identificar si un Tweet perteneciente a cada categoría correspondiente es posible responderlo de forma automatizada utilizando las respuestas creadas anteriormente.

7.4.- Medidas de Evaluación

Una vez finalizada el proceso de respuestas automáticas es importante medir que tan preciso son las reglas generalas que se han creado, para lograr esto se utilizaran las métricas utilizadas anteriormente en el proceso de clasificación de Tweets.

7.4.1.- Respuesta de Agradecimiento

Tabla 10.- Resumen Métricas Respuesta Agradecimiento

Total Categoría	Total Respuestas Clasificadas	Total Respuestas Automáticas	Precisión	Accuracy	Recall	F1-Score
1173	548	493	97%	92%	87%	91%

	Clasificación	
Clase	Si	No
Respuesta	476	72
No Respuesta	17	608

Figura 33.- Matriz Confusión Respuesta Agradecimiento

7.4.2.- Respuesta de URL

Tabla 11.- Resumen Métricas Respuesta URL

Total URL	Total Respuestas Clasificadas	Total Respuestas Automáticas	Precisión	Accuracy	Recall	F1-Score
591	172	188	85%	93%	93%	89%

	Clasificación	
Clase	Si	No
Respuesta	160	12
No Respuesta	28	391

Figura 34.- Matriz Confusión Respuesta URL

7.4.3.- Respuesta Alumbrado Público

Tabla 12.- Resumen Métricas Respuesta Alumbrado Público

Total Alumbrado Público	Total Respuestas Clasificadas	Total Respuestas Automáticas	Precisión	Accuracy	Recall	F1-Score
2279	27	22	91%	100%	74%	82%

	Clasificación	
Clase	Si	No
Respuesta	20	7
No Respuesta	2	2250

Figura 35.- Matriz Confusión Respuesta Alumbrado Público

7.4.4.- Respuesta DM

Tabla 13.- Resumen Métricas Respuesta DM

Total DM	Total Respuestas Clasificadas	Total Respuestas Automáticas	Precisión	Accuracy	Recall	F1-Score
2279	19	19	95%	100%	95%	95%

	Clasificación	
Clase	Si	No
Respuesta	18	1
No Respuesta	1	2259

Figura 36.- Matriz Confusión Respuesta DM

7.4.5.- Respuesta Consultas Generales

Tabla 14.- Resumen Métricas Respuesta Consulta Generales

Total Consultas Generales	Total Respuestas Clasificadas	Total Respuestas Automáticas	Precisión	Accuracy	Recall	F1-Score
2279	103	81	74%	97%	58%	65%

	Clasificación	
Clase	Si	No
Respuesta	60	43
No Respuesta	21	2155

Figura 37.- Matriz Confusión Respuesta Consultas Generales

7.4.6.- Respuesta Numero Emergencias

Tabla 15.- Resumen Métricas Respuesta Numero Emergencias

Total Consultas Generales	Total Respuestas Clasificadas	Total Respuestas Automáticas	Precisión	Accuracy	Recall	F1-Score
2279	31	28	96%	100%	87%	92%

	Clasificación	
Clase	Si	No
Respuesta	27	4
No Respuesta	1	2247

Figura 38.- Matriz Confusión Respuesta Numero Emergencias

7.4.7.- Respuesta Pregunta Operatividad

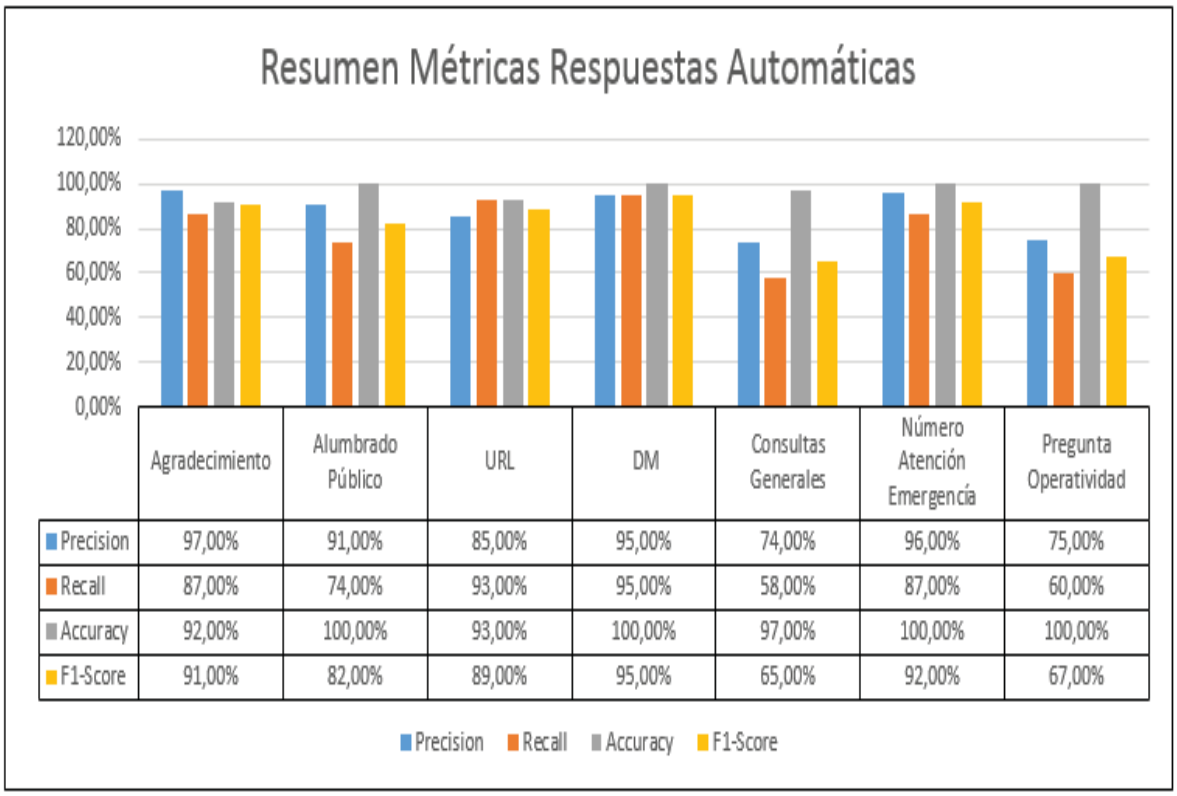
Tabla 16.- Resumen Métricas Respuesta Operatividad

Total Pregunta Operatividad	Total Respuestas Clasificadas	Total Respuestas Automáticas	Precisión	Accuracy	Recall	F1-Score
2279	5	4	75%	100%	60%	67%

	Clasificación	
Clase	Si	No
Respuesta	3	2
No Respuesta	1	2273

Figura 39.- Matriz Confusión Respuesta Operatividad

A continuación se mostrara un gráfico que resumirá los datos mostrados anteriormente.



	Agradecimiento	Alumbrado Público	URL	DM	Consultas Generales	Número Atención Emergencia	Pregunta Operatividad
Precision	97,00%	91,00%	85,00%	95,00%	74,00%	96,00%	75,00%
Recall	87,00%	74,00%	93,00%	95,00%	58,00%	87,00%	60,00%
Accuracy	92,00%	100,00%	93,00%	100,00%	97,00%	100,00%	100,00%
F1-Score	91,00%	82,00%	89,00%	95,00%	65,00%	92,00%	67,00%

Figura 40.- Resumen Métricas Respuestas Automáticas

8.- Conclusiones y Trabajos Futuros

8.1.- Conclusiones

En el presente se ha optado por utilizar cuatro representaciones las cuales corresponden a la representación TF-IDF, Binaria y la representación TF-RFL, además se ha implementado el método de Validación Cruzada, el cual ha servido para distinguir cual método de los utilizados tiene un mejor desempeño al momento de clasificar dando como resultado el método de Arboles de Decisión con la representación TF-RFL, permitiendo clasificar de buena forma un gran porcentaje de los Tweets. A demás el método de Maquinas de Soporte Vectorial también a logrado resultados significativos mediante la representación Binaria.

Otro importante resultado que cabe destacar es el uso de una representación no tan común como la TF-RFL que ha arrojado buenos resultados 97% de Accuracy con el método de Árbol de Clasificación, logrando además el mejor resultado en cuanto a la precisión de las tres representación usadas con un valor de 97,4%, una conclusión importante que se puede resaltar es la comparativa que se ha obtenido al realizar el proceso con la representación TF-RFL y esta misma pero con los valores normalizados, dado los resultados mostrados en las versiones preliminares de este trabajo, no hay variación importante entre realizar la normalización a esta representación.

Cabe señalar el buen comportamiento que se ha logrado a través de las reglas generales para clasificar de forma correcta las respuestas automáticas dándose muy buenos resultados en algunos casos como por ejemplo la respuesta a los Tweets de agradecimiento, la cual logro una precisión del 97%, sin embargo hay algunas respuestas que no se obtuvieron buenos resultados esto se puede deber a la gran variedad de casos que abarca la respuesta. Esto puede provocar molestias, ya que, se responden Tweets que no debieran ser respondidos y esto provocara que los usuarios no se sientan escuchados generando malestar en estos, repercutiendo en una mala imagen para la empresa. Esto da pie para pulir más las reglas generales y buscar mejores resultados.

También es importante señalar la importancia del proceso de reducción de características, ya que, gracias a él se ha disminuido enormemente el Corpus generado ya que en una primera instancia se había generado un Corpus de 7860 palabras de las cuales una gran cantidad no aportaba información para clasificar a un Tweet en algunas de las categorías, quedando finalmente de un tamaño de 2845 palabras que en vista de los resultados obtenidos se puede decir que es bastante representativo.

8.2.- Trabajos Futuros

En cuanto al trabajo futuro de las respuestas automáticas es posible mejorar aún más los resultados obtenidos agregando nuevas palabras representativas al corpus de cada respuesta y generando nuevas reglas que permitan identificar de mejor manera si es posible responder de forma automatizada un Tweet.

9.- Referencias

[1] Social Media Benchmark Study, 2013

[2] Baeza-Yates & Ribeiro-Neto Berthier, Modern Information Retrieval. Addison-wesley, Wokingham, UK, 1999

[3] Sebastiani, Fabrizio, Machine learning in automated text categorization. ACM Computing Surveys, Vol.34 Nº.1, pp 1-47, 2002

[4] Kagan Tumer y Joydeep Ghosh, "Order Statistics Combiners for Neural Classifiers," World Congress on Neural Networks, 1995.

[5] Hrvoje Bacan, Igor S. Pandzic y Darko Gulija, "Automated News Item Categorization," 2005.

[6] Figuerola, Alonso Berrocal, Zazo Rodríguez, & Rodriguez, La recuperación de información en español y la normalización de términos 2004

[7] Araujo N, Método Semisupervisado para la Clasificación Automática de Textos de Opinión. 2009

[8] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up: sentiment classication using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79{86, Association for Computational Linguistics, 2002.

[9] Ass K. & Eikvil L. "Text categorization: A survey". Technical Report. Norwegian Computing Center, 1999.

[10] Alfaro R., Allende H. (2010), Text Representation in Multi-label Classification: Two New Input Representations 10th International Conference on Adaptive and Natural Computing Algorithms (ICANNGA'11).