# UNIVERSIDAD VERACRUZANA

## CENTRO DE INVESTIGACIONES EN INTELIGENCIA ARTIFICIAL

# MÉTODO DE PREDICCIÓN DE TENDENCIAS EN TWITTER BASADO EN APRENDIZAJE INCREMENTAL SOBRE UNA PLATAFORMA MULTI-AGENTES

TRABAJO RECEPCIONAL EN LA MODALIDAD DE:

**TESIS**

COMO REQUISITO PARCIAL PARA OBTENER EL TÍTULO DE

**DOCTOR EN INTELIGENCIA ARTIFICIAL**

PRESENTA:

**M.I.A. YECELY ARIDAÍ DÍAZ BERISTAIN**

DIRECTOR

DR. GUILLERMO DE JESÚS HOYOS RIVERA

CODIRECTOR

DR. NICANDRO CRUZ RAMÍREZ

VERACRUZ, VER. ENERO 2018

# ÍNDICE

# RESUMEN

En el presente trabajo se presenta la conceptualización y puesta en operación de un método de predicción de tendencias en Twitter, el cual hace uso de un algoritmo de aprendizaje incremental conocido como Árbol de Hoeffding, y cuya implementación se llevó a cabo siguiendo un modelo de arquitectura de Sistemas Multi-Agentes BDI. La investigación tuvo como objetivo: 1) determinar los factores ocultos que ayudan a posicionar un hashtag en Twitter, 2) presentar una arquitectura basada en Sistemas Multi-Agentes que aporta autonomía en el método, 3) con base en los elementos anteriores, generar posibles reglas de asociación, basadas en la clasificación incremental mediante el Árbol de Hoeffding, que den como resultado una mayor oportunidad de buen posicionamiento de un hashtag en función del tema del que se esté hablando, y 4) evaluar y comparar los resultados con métodos de predicción ya existentes.

# CAPÍTULO I: Introducción

Desde la aparición de la Internet, han surgido diversas herramientas que habilitan la comunicación entre usuarios, como lo ha sido el correo electrónico, Gopher, Usenet News, etc. La propuesta que más relevancia a tomado con el paso del tiempo es la de la llamada World-Wide Web, o simplemente Web, y que ahora domina ampliamente el uso de los recursos de la Internet. La Web surge, en 1989, a partir de una propuesta de Sir Timothy “Tim” John Berners-Lee al interior del CERN (Conseil Européen pour la Recherche Nucléaire - Consejo Europeo para la Investigación Nuclear), en Suiza, como una herramienta que le permitiera a él, y a sus colegas, intercambiar fácilmente información. Su arquitectura era muy simple, y lo sigue siendo, y consiste de un servidor y uno o más clientes.

Poco a poco la propuesta inicial fue ganando relevancia, hasta que se convirtió en un éxito de uso a nivel mundial. Sin embargo, durante muchos años el modo de operación siguió siendo el mismo, bajo un paradigma en el cual sólo había pocos editores de contenido, conocidos como Web Masters, y el resto eran simplemente consumidores de éste.

La denominada Web 2.0 [1] surge a partir de un proceso evolutivo en el que la Web (a la que ahora se puede llamar 1.0) cambia radicalmente su modo de operación a partir de la integración de herramientas que habilitan a los usuarios para interactuar entre ellos, y convertirse en actores, en vez de simples espectadores. Ésta consiste de aquellos sitios Web que dotan al usuario de las capacidades de compartir información, colaborar e interactuar con otros usuarios. El surgimiento de la Web 2.0 logró que el usuario desempeñe un rol preponderante, y cambia de ser un simple espectador, a ser un ente activo capaz de influir, a través de sus intervenciones, en la forma de pensar de otros usuarios. “La Web 2.0 es participativa por naturaleza. En ella, los usuarios no suelen adoptar una actitud pasiva, sino todo lo contrario. No sólo

leen, también discuten, comentan, escriben, publican, intercambian, escogen, corrigen, comparten, es decir, participan activamente” [2].

De esta manera el conocimiento se va creando con la aportación de los usuarios mediante el uso de blogs [3, 4], videos [5, 6], Wikis [7], sitios colaborativos [8, 9], Redes Sociales, etc. Este proceso evolutivo en la Web genera información relevante a través de recursos que, poco a poco, se han ido generando de manera más o menos natural, como parte del mismo proceso evolutivo. Tal es el caso de las meta-etiquetas, los hashtags, las menciones, etc., y en función de la plataforma de la que se esté haciendo uso, logrando así que los datos tengan un valor agregado, producto de la diversidad de la manera de pensar de los usuarios que las usan. Entre los sitios Web más destacados que pueden ser catalogados como pertenecientes a la Web 2.0 se encuentran: los videos (Youtube® y Vimeo®), wikis (Wikipedia®, Wikilengua®) y RS (Instagram®, Facebook® y Twitter®), solo por mencionar algunos.

Una de las aportaciones más notables de la Web 2.0 son las Redes Sociales Basadas en la Web, o simplemente llamadas, de manera genérica, Redes Sociales (RS en lo subsecuente). Éstas son una extensión digital del concepto mismo de red social utilizado desde hace ya varios años en Sociología. Cada red social de esta naturaleza puede ser definida como “una gran red social interconectada que recopila información sobre los contactos sociales de los usuarios” [10]. Sus inicios se pueden explicar mediante la teoría del “mundo pequeño”, postulada por Stanley Milgram y Jeffrey Travers, en 1969, que parte de la afirmación de que, para un grafo en el que la mayoría de los nodos no son vecinos, es posible llegar a la mayoría de esos nodos desde cualquier otro nodo a través de las relaciones con un conjunto de nodos intermedios. Diversos investigadores se han interesado en conocer la forma en que las personas tienen conexiones entre ellas aunque estén en diferentes ciudades o solo mediante amigos en común [11, 12]. Los autores de [13] definen a la RS como “un servicio que permite a las personas construir un perfil público o semipúblico dentro de un sistema limitado, organizar la lista de usuarios que comparten una conexión y conectar unos con otros”. Ellos mencionan que la naturaleza y

nomenclatura de las conexiones pueden varias de una RS a otra. Las RS pueden influir en el comportamiento de un individuo, pero también reflejan sus preferencias, intereses y opiniones.

Poco a poco las RS se están convirtiendo en un importante medio de comunicación que genera diariamente una vasta cantidad de datos a partir de las acciones e interacciones de sus usuarios, como es el caso de las publicaciones que realizan, su historial de navegación, las conexiones que tienen con otros usuarios, el uso de emojis, el uso de contenido multimedia, etc. Además se han convertido en una herramienta de comunicación cada vez más importante en diversos ámbitos como política, salud, entretenimiento, mercadotecnia, ciencias sociales, etc. Esto da como resultado la aparición de un área de oportunidad en términos del análisis de datos con la finalidad de contribuir al diseño de modelos y métodos permitan explotar, de alguna manera, toda la información asociada a estas actividades.

Actualmente, las RS son ampliamente conocidas alrededor del mundo. Sin embargo su proceso evolutivo, si bien vertiginoso, ha seguido varias etapas. En 1997 se vio el nacimiento de una de las primeras RS con el surgimiento de Six Degrees® [14] la cual fue utilizada brevemente entre 1997 y 2001, llegando a alojar a 3.5 millones de usuarios registrados en su apogeo. Su nombre lo recibió a partir de la teoría de los seis grados de separación. Esta teoría intenta comprobar la hipótesis de que cualquier persona en una población dada puede estar conectada con cualquier otra a través de una cadena que no tiene más de cinco intermediarios, conectando a ambas personas con solo seis enlaces. El sociólogo Duncan Watts, quien es quien postuló esta teoría, asegura que es posible lograr esta relación y lo plantea en su libro “Seis grados: la ciencia de una edad conectada” [15].

Algunos años después de Six Degrees nace Friendster®, creada por un programador canadiense llamado Jonathan Abrams en 2002, logrando más de 5 millones de usuarios registrados en el lapso de algunos meses. Desafortunadamente comenzó a experimentar dificultades técnicas al punto en que los usuarios decidieron

abandonarla buscando otras alternativas. Chris De Wolfe, Tom Anderson, Josh Berman y Brad Greenspan se percataron del potencial y fundaron MySpace en octubre de 2003. MySpace fue lanzada oficialmente en enero de 2004 y en su primer mes un millón de personas se registraron, su objetivo inicial se centró en poder conectar a diferentes usuarios y principalmente en compartir contenido multimedia. En 2006, fue nombrado como el sitio más visitado de los Estados Unidos, y valorado en 2007 en decenas de miles de millones de dólares cuando alcanzó su pico máximo de visitas. Su problema principal fue la mala administración de sus fundadores, perdiendo usuarios por la falta de visión y poca o nula mejora en sus servicios. Tales RS permitieron que los usuarios tuvieran una manera de comunicarse, conectarse y compartir información con amigos y personas con intereses comunes. Fue el inicio de una nueva etapa en la comunicación.

Posteriormente han existido diversos intentos de creación de herramientas para el manejo de RS, sin embargo la mayoría de ellas han pasado sin dejar huella o legado de relevancia.

Actualmente, si bien es cierto que existe toda una pléyade de RS en el mundo de Internet, las dos que más importancia han ganado son Facebook y Twitter. Facebook es la RS más utilizada en todo el mundo, con aproximadamente dos mil millones de usuarios activos, mientras que Twitter tiene 328 millones de usuarios activos. Ambas se destacan por su intensidad de uso, número de usuarios activos, e incremento en el tráfico de los datos, convirtiéndose en un medio donde los usuarios interactúan y pueden establecer relaciones con otros usuarios con intereses comunes.

Si bien es cierto que Twitter tiene menos usuarios alrededor del mundo que Facebook, estos tienen la característica de ser más activos en la primera que en la segunda. Otra diferencia sustancial es el modelo que sigue cada una de estas plataformas para efecto del establecimiento de las relaciones entre sus usuarios. En el caso de Facebook, es un modelo simétrico, en el que, para poder establecer una relación entre dos usuarios, es indispensable que ambos accedan a tal hecho,

mientras que en Twitter el modelo es asimétrico, por lo que no es necesaria una autorización para el establecimiento de una relación.

Las principales razones que han ejercido influencia para seleccionar a Twitter como RS para la realización de nuestro estudio son:

1. Es una RS abierta que no requiere de una relación con todos los usuarios que participan en el tema para obtener sus tweets publicados, únicamente es necesario usar la API de Twitter para la extracción de las publicaciones de interés;
2. El contenido de los mensajes es corto (140 caracteres[1]), lo que permite que la extracción de datos y la aplicación de técnicas de minería de datos sea menos compleja;
3. La agrupación de los tweets por HTs favorece el estudio e interpretación de casos por separado, encontrando correlaciones entre temas y contenido.

La característica distintiva de Twitter de limitar los mensajes a una longitud de 140 caracteres como máximo hace que esta RS pueda ser considerada como un servicio de micro-blogging. Los mensajes publicados en esta plataforma son comúnmente conocidos como tweets y además de texto pueden incluir imágenes estáticas, encuestas, GIFs, URLs y videos. Twitter es una RS de información en tiempo real sustentada por los mensajes publicados por personas de todo el mundo, actualmente se publican alrededor de 500 millones de tweets por día.

Debido a la enorme cantidad de datos propagados en esta red, los usuarios han desarrollado diferentes medios que les permiten facilitar su interacción y comunicación, y en cierta forma, organizarla. El caso más emblemático es el uso de los denominados *hashtags*. Un hashtag (HT) es una palabra, o secuencia de palabras concatenadas, antecedidas del prefijo numeral (# - hash en inglés). A través de éstos se pueden etiquetar o agrupar tweets asociados a un tema específico,

[1] Actualmente la longitud ya es de 280 caracteres, pero es un cambio que se implementó una vez ya concluida la fase de investigación e implementación del presente trabajo.

identificado a través del HT mismo, y los usuarios pueden dar seguimiento al flujo de cualquier conversación. Algunos ejemplos de HT son: #TodosSomosUV o #UVDamosMas. Los HTs tuvieron su origen a inicios de 1993 como parte del sistema de mensajería IRC (Internet Relay Chat), ya en desuso, utilizándose para crear canales dedicados a temas específicos donde los usuarios se integraban para hablar de algún tema específico. En 2007, un desarrollador de Google llamado Chriss Messina propuso al HT como una solución para relacionar los mensajes de los usuarios en Twitter, y posteriormente a Nate Ritter se le conoce como la primer persona en hacer efectivamente uso de un HT: #SanDiegoFire, para referirse a los mensajes sobre los incendios forestales en California. Una característica importante implementada a partir del 1º de Julio de 2009, es que Twitter añade un enlace a cada HT, de forma que a través de éste es posible visualizar los tweets que lo incluyen en su contenido, y agruparlos por: destacados, más reciente, personas, fotos, videos, noticias y transmisiones. Entre los HT más relevantes a lo largo de la historia de Twitter se encuentran #loveTwitter utilizado para festejar los 10 años de la RS, #PrayForParis relacionado con los ataques terroristas en París el 13 de noviembre de 2015, #Mandela para comentar sobre el fallecimiento del ex-presidente de Sudáfrica Nelson Mandela, y que inundó Twitter con un aproximado de 90,000 tweets por minuto, sólo por mencionar algunos.

Adicionalmente, los usuarios de Twitter pueden hacer uso de otro recurso: la mención. Las menciones se caracterizan por utilizar el prefijo arroba (@) y posteriormente el nombre del usuario al que se desea referir, de manera que se le indicará al usuario mencionado que hay un comentario o publicación donde se hace referencia a ésta, y que, por ende, requiere de su atención. Cualquier usuario puede mencionar a cualquier otro usuario, a menos que el primero de ellos haya sido explícitamente bloqueado por el segundo.

Los tweets publicados por los usuarios son colocados en la, así llamada, línea de tiempo (TL) ó *timeline* en inglés. Ésta consiste en una cronología de los tweets publicados por un usuario dado. El *timeline* de un usuario puede ser visto

simplemente accediendo a su perfil, donde son mostrados de manera cronológica del más reciente al más antiguo, y en dado caso que el usuario haya seleccionado un tweet como *fijado,* entonces se muestra primero sin importar la fecha en la que fue publicado.

Finalmente, los usuarios de Twitter pueden establecer relaciones entre ellos. Estas relaciones son, como se mencionó anteriormente, asimétricas, pues no requieren de una autorización para que puedan ser creadas, y son denominadas “seguir”, o *Follow* en inglés. Cualquier usuario de Twitter podrá seguir a cualquier otro usuario sin más protocolo que aplicar la acción de seguimiento a través de cualquiera de las implementaciones de la interfaz de Twitter. A partir de ese momento, cualquier tweet publicado por el segundo, será automáticamente presentado en la pantalla principal de Twitter del primero tan sólo sea emitido.

Una de las acciones mas comunes en Twitter, aparte de la de publicar tweets, es la de difundir, o retransmitir, algún tweet previamente publicado por algún otro usuario, o a su vez retransmitido por este último. A esta acción se le denomina “hacer retweet” (RT), y es la que distingue más claramente a esta plataforma. Lo que caracteriza a los RT es que el texto original, incluido cualquier eventual HT dentro de él, es vuelto a publicar dentro de la plataforma por otro usuario, lo que respeta el espíritu original del mensaje. Esta peculiaridad hace que ciertos mensajes tengan un comportamiento similar al de las infecciones biológicas, a una pandemia o epidemia. La infección, en este contexto, se hace a través de la mente de los usuarios, cuando estos son “tocados” por un mensaje que los hace reflexionar y reaccionar, y que los motiva a hacerlo propio a través la retransmisión.

De esta forma nace el concepto de viralidad dentro de Twitter, haciendo un símil con el comportamiento de los virus computacionales, que a su vez fueron llamados así por la similitud de su proceder con las infecciones de naturaleza biológica. La viralidad no puede existir si el tweet pertenece exclusivamente a un usuario, para la supervivencia y difusión necesita extenderse en la RS, esto mediante el RT.

El comportamiento de los tweets en términos de su difusión, suele ser más o menos estable, sin embargo hay ocasiones en las que algún usuario crea un tweet con cuyo tema se identifica un importante número de usuarios de la plataforma, razón por la cual lo difunden (retuitean, según el argot comúnmente usado), pudiendo llegar a un número de RT muy grande. En tales casos, cuando la cantidad de RT llega a ser desmedido que sobrepasa ciertos límites en un periodo dado, establecido por Twitter, se convierte en un tema tendencia, o *Trending Topic* (TT) en inglés.

Establecidos los hechos anteriormente descritos, y pasando al tema que nos atañe, podemos afirmar que la presente investigación se enfoca en tres elementos principales: los *hashtags* (HT), *retweets* (RT) y *trending topics* (TT). Debido a la cantidad de información, así como el interés por comprender el comportamiento de los usuarios, opiniones, análisis de sentimiento en el texto, entre otros elementos de información, los tweets son un medio importante para lograrlo. Sin embargo, para que las propuestas sean realmente útiles es necesario recopilar de manera apropiada los tweets y los elementos que lo acompañan. Una parte clave del proceso de investigación se logra con el uso de técnicas de aprendizaje automático que sean adecuadas para limpiar el ruido en la datos y seleccionar los atributos clave para el análisis y aprendizaje. Cuando se recopilan los tweets en las RS en general, en ocasiones se utilizan palabras clave que facilitan la extracción de la información. En el caso de Twitter el uso de los HT simplifican la extracción, siendo uno de los elementos fundamentales por los cuales el caso de estudio se centra en esta RS. Además, la longitud limitada a 140 caracteres da como resultado un tratamiento más simple de los datos.

Para realizar el proceso de extracción de los tweets hacemos uso de la Interfaz de Programación de Aplicaciones (API) de Twitter, a través de la cual se puede solicitar la recuperación de cualquier tipo de información relativa a la operación de la misma RS. Es a partir de la información recuperada que estamos en posibilidad de llevar a cabo diversos análisis del presente trabajo.

Este documento de tesis está organizado de la siguiente manera: en el capítulo II presentamos el Estado del Arte y los trabajos relacionados con la presente investigación. El capítulo III se enfoca en la metodología y los elementos que la componen. En el capítulo IV presentamos la arquitectura basada en Sistemas Multi-Agentes de tipo BDI. Finalmente en el capítulo V los experimentos y resultados del método propuesto, y el capítulo VI presenta las conclusiones y trabajo a futuro.

# CAPÍTULO II: Estado del Arte

Twitter, como se mencionó en el capítulo anterior es una RS en constante crecimiento y cuyos usuarios tienen la característica de ser muy activos, y en dónde pueden compartir sus ideas, pensamientos, intereses y opiniones en mensajes cortos de 140 caracteres como máximo llamados tweets. Con un aproximado de 328 millones de usuarios activos y 500 millones de tweets por día se convierte en una fuente de información invaluable. Por ello, el estudio de Twitter se ha extendido más allá de comprender los intereses en común que colaboran al seguimiento entre usuarios [16, 17].

Actualmente hay distintos aspectos de la RS que han sido investigados de manera profunda, como es la popularidad de los usuarios [18], el análisis de sentimiento en los tweets [19, 20] y la extracción de la información [21, 22], etc. El objetivo de esta investigación se enfoca primordialmente en presentar un método de predicción de TT en Twitter basado en un enfoque incremental mediante Sistemas Multi-Agentes, siendo el comportamiento de los HT una pieza clave de la investigación realizada. Para poder predecir el crecimiento de un HT o tema, es necesario primero, comprender aspectos básicos, y responder a preguntas como:

- ¿Por qué los usuarios se conectan?
- ¿Que hace a un tweet popular?
- ¿Por qué un usuario hace un RT?
- ¿Por qué las personas hablan de ciertos temas y otros los ignoran?.

Estas preguntas son algunas de las interrogantes que investigadores han tratado de responder, y que son tomadas como base para generar nuevo conocimiento.

En este capítulo presentamos los resultados de las investigaciones realizadas por otros equipos de trabajo abordando este mismo tema, y partiendo de éstas, procedemos a posicionar nuestra propuesta.

En la primera etapa se examina el comportamiento de los HT y las propuestas de diversos autores en este sentido, para posteriormente analizar las investigaciones relacionadas con el estudio de los TT y por último investigaciones que utilizan clasificadores incrementales.

### *2.1. Comportamiento de los HT*

Los HT han logrado un importante interés de investigación especialmente por ser una forma de difusión agrupada de acuerdo a un tópico. Sin embargo, aún no se tienen algoritmos con un nivel de precisión suficientemente alto que permitan categorizarlos inequívocamente. Por medio de los HT, y las estructuras formadas de quienes los integran en sus tweets, es posible comprender el TT de los temas en la RS y que grupos se están formando, entre otros datos.

Una de las propuestas que se enfoca en que los HT sean categorizados, y a partir de ahí descubrir patrones de interés de los usuarios, es la realizada por Romero et al [23]. En ésta se proponen ocho categorías a partir del análisis de 500 HT, entre las que se encuentran: celebridades, música, películas y televisión, videojuegos, política, idiomas, deportes y tecnología. Para lograrlo extrajeron mediante la API de Twitter un historial de tweets del periodo comprendido entre agosto de 2009 a enero de 2010, además de extraer los tweets con sus respectivos HT también construyeron la red de usuarios. Conectaron al usuario *X* con el usuario *Y* siempre y cuando *X* mencionó al menos 3 veces a *Y*. Como resultado crearon una red que contiene 8.5 millones de nodos aislados y 50 millones de enlaces. Concluyeron que los HT se propagan mediante exposiciones repetidas que proporcionan un principio de “contagio”, de manera que mientras más se hable de un tópico es probable que nuevos usuarios se unan a la conversación principalmente cuando se tienen intereses en común, o mejor conocida como “homofilia”.

El término homofilia se refiere al hecho de que “el contacto entre personas similares ocurre a un ritmo mayor que entre personas diferentes” [24]. Los autores mencionan que la similitud genera conexión y que puede lograrse por medio de vínculos de todo tipo que incluyen la amistad, trabajo, religión, nivel socio-económico, entre otras.

Por otro lado, Cunha et al. [25] analizaron la evolución del HT y el interés de los usuarios al crearlos, utilizarlos y difundirlos con el objetivo de comprender mejor su propagación. Para ello analizaron un total de 1.7 millones de tweets correspondientes al periodo entre Julio de 2006 y Agosto de 2009, de manera que se enfocaron a buscar características en los mensajes y posibles patrones entre los diferentes HT. Debido a la cantidad de mensajes se enfocaron exclusivamente en estudiar algunos temas, como son: la muerte de Michael Jackson, la epidemia H1N1 y un evento llamado “Music Monday”. Concluyen que, debido a que su enfoque es basado en técnicas de lingüística, su análisis mostró similitudes cualitativas y cuantitativas, revelando que existe una correlación entre el uso de palabras simples del lenguaje y su propagación. Para lograr dichos resultados utilizaron la Ley de Zipf [26] que consiste en determinar la frecuencia de aparición de distintas palabras en un texto. Determinan que los HT más populares son los más simples, directos y cortos, comparados con aquellos que incluyen cadenas largas de caracteres y además palabras complejas.

Weng et al. [27] orientaron su investigación a los intereses que comparten los usuarios, y cómo ésta afecta en el crecimiento de los HT. Para ello consideraron dos contextos: ubicación geográfica y población. Fue importante que los usuarios tuvieran su geolocalización activada y cuando no contaron con esos datos, asignaron un indicador aproximado de acuerdo a su zona horaria. A partir de ello, calcularon la diferencias de tiempo entre los usuarios que se comunicaron a través de los tweets publicados. Al finalizar la investigación concluyeron que los usuarios buscan relacionarse con otros que les interesan temas similares, y que hay usuarios que son particularmente selectivos sobre su decisión de a quién seguir.

En consecuencia podemos inferir que, si deseamos que un tópico sea ampliamente compartido, se debe encontrar a ese grupo de usuarios que demuestran un interés con el tema y llegar a sus seguidores de manera que se establezca una atracción por las publicaciones.

### *2.2. Tendencias en Twitter*

Twitter permite observar los TT fácilmente y saber cuáles son los temas de interés para las masas digitales, ya sea en un país específico, o de manera global, teniendo como alimento flujos de datos en tiempo real. Kwaw et al. [28] muestran que diferentes tweets que han llegado a ser TT, a pesar de que comparten la peculiaridad de ser populares, tienen características específicas diferentes en términos de cantidad de respuestas, menciones, RT y contenido. Su conclusión fue que la mayoría (85%) de los temas observados son noticias del momento o son estacionales. En el caso del trabajo motivo del presente documento, el 65% de los temas mencionados tuvieron una relación con otros medios digitales y los restantes surgieron en la misma RS como son: #SaboteaUnWalmart, #MientoComoArne o #BroncoVSPeje. También los autores mencionan que a partir de un promedio de 1,000 usuarios que hablan del tema sin importar su número de seguidores, otros usuarios se incluirán en la conversación. De acuerdo a las observaciones realizadas en la investigación, nosotros no coincidimos con estos resultados, puesto que detectamos que es importante primero llegar a usuarios con un mayor número de seguidores, aunque la muestra sea menor, ya que es más sencilla la propagación.

Otro enfoque es el de Sankaranarayanan et al. [29], quienes usan técnicas de agrupamiento para identificar los TT y clasificar los tweets según su categoría. El objetivo principal de los autores es clasificarlos de manera automatizada y no manual como lo realizado por Kwaw et al., y principalmente lograr un método de clasificación con una precisión alta y que pueda ser entrenado por medio del histórico de los tweets.

Zubiaga et al. [30] también proponen una manera de clasificar los TT. Definen una topología basada en 4 categorías: *noticias, eventos en curso, memes y conmemorativos*. La primera se refiere a los eventos periodísticos o televisivos que se producen en las últimas horas y que empiezan a difundirse en Twitter, posteriormente los eventos en curso son aquellos que relatan hechos en tiempo real, los cuales pueden ser un juego de futbol, presentación de tecnología, festival de música, entre otros. La categoría *memes* son aquellos mensajes que pueden ser divertidos y que pueden incluir o no una imagen. De acuerdo con los autores es la más difícil de clasificar debido a que también llegaron a considerar imágenes relacionados a alguna protesta. Y por último, los conmemorativos, son aquellos en que los usuarios publican mensajes celebrando alguna fecha como la navidad. Para realizar una propuesta de clasificación automática, diseñaron diferentes diccionarios relacionados con cada categoría, de manera que mediante el número de ocurrencias dentro de cada mensaje le asignaron su etiqueta correspondiente. Además utilizaron Máquinas de Soporte Vectorial para considerar la asignación de multi-etiquetas. Concluyen que el uso de algoritmos alternativos podría ayudar a cuantificar la asignación de las etiquetas a cada TT.

Naaman et al. [31] también proponen caracterizar los TT emergentes en dos grupos principales denominados exógeno y endógeno. La primera se refiere a actividades, intereses y eventos que ocurren fuera de Twitter, y la segunda categoría, todos los mensajes que son actividades realizadas dentro de la RS. A partir de estas dos categorías, se definen subcategorías, como emisión de multimedia, eventos globales, fechas de celebración, *memes*, RT y actividades de comunidades. Una vez definidas las categorías, y asignadas a un grupo de mensajes extraídos, concluyeron que los TT exógenos tienen una menor proporción de RT que los TT endógenos. Además, logran una mayor conversación entre los usuarios de la RS. Por otro lado, los tópicos clasificados como *memes* tienen menor interacción que las restantes. Concluyen que sus resultados son cuantitativos y ofrecen indicadores que pueden utilizarse para caracterizar con mayor precisión a los TT.

Recuero et al. [32] proponen una estrategia diferente para estudiar los TT, y es a través del comportamiento y aplicación de los usuarios. Ellos sugieren utilizar un enfoque cualitativo y cuantitativo que describa el comportamiento de los TT. Para ello se enfocaron a un caso de estudio en particular: el HT creado para una banda de rock brasileña llamada “Restart”. Extrajeron los mensajes que incluían #Restart en su contenido y a los usuarios que lo utilizaron, posteriormente analizaron y crearon un grafo de conexiones mediante la herramienta NodeXL. En su caso de estudio notaron si el posicionamiento del HT es realizado por la noche tienen más probabilidades de qué se convierta en un TT. Por último, incluyeron en el mensaje HT que estaban posicionados como TT y de esta manera llegaron a otros grupos de conversación más populares. Aunque su propuesta es enfocada a un experimento empírico, nos permite ratificar algunas de nuestras hipótesis posteriormente planteadas.

En la presente investigación se muestran los atributos que comparten los HT analizados, ya sean los TT, o aquellos que pasaron desapercibidos, por lo que se tiene una aproximación de aquellos atributos de interés para los usuarios, y que pueden utilizarse posteriormente para popularizar un tema.

### *2.3. Métodos de predicción*

En el caso de Asur et al. [33], explican la popularidad de los tópicos en Twitter, y determinan los factores que contribuyen a la creación y evolución de los TT enfocándose principalmente en la persistencia del tópico. Su estudio se basa en la frecuencia e intervalos de los tweets desde el momento en que comienza a mostrarse una tendencia hasta que desaparece. A partir de esto obtienen un número acumulado de tweets y el tiempo en que le tomó ser popular. Los autores mencionan que es normal que los temas sean divulgados con mayor frecuencia al inicio y que frenen su crecimiento a medida en que se vuelven obsoletos de poco interés para los usuarios que se unen a otros HT. De ésta manera el número de mensajes disminuye con el tiempo. Es por ello que mientras más usuarios publiquen del tema, es más

probable que otros usuarios se unan a la difusión del tema. Otros autores coinciden en que es importante llegar a aquellos usuarios denominados como influyentes, de forma que, si ellos hablan del tema o dan un RT, es más probable que existan picos de distribución de los mensajes. Los autores describen y comprueban sus hipótesis antes mencionadas, sin embargo, no determinan factores individuales del tweet, se enfocan principalmente en los grupos y dan por hecho que el HT está en las primeras posiciones. En nuestro caso partimos de las características necesarias de los tweets de manera individual y en conjunto, de forma que los atributos más allá de los evidentes puedan ser aplicados para incrementar las posibilidades de creación de una TT.

Zaman et al. [34] se enfocan específicamente en los tweets y RT y mediante una metodología se proponen predecir si un usuario hará un RT. De la misma manera que otros autores se enfocaron inicialmente en la obtención de un conjunto de tweets mediante el uso de la API de Twitter, con el fin de entrenar modelos, siendo en este caso probabilísticos de filtrado colaborativo. Sus modelos aprendieron patrones considerando tres criterios: (i) el usuario que publica, (ii) el usuario que hace el RT y (iii) el contenido del tweet. Descubrieron que la identidad de la fuente del tweet y el usuario que da el RT son las características más efectivas para predecir los RT futuros.

Su metodología se basa en el modelo Matchbox, que fue diseñado originalmente para predecir la preferencia de los usuarios por cierto tipo de películas, siendo los metadatos su fuente principal. En este caso, las entradas al modelo son el usuario que publicó, el usuario que hizo RT y el contenido del tweet, y como salida se obtiene una distribución de probabilidad, que refiere a la probabilidad de que un tweet obtenga RT por cierta cantidad de usuarios.

El objetivo de la investigación es aprender de las correlaciones entre los usuarios que hacen RT y la preferencia del tweet. Si algún usuario ha hecho RT al mensaje publicado entonces se obtiene el valor de 1, si el tweet en un periodo de una hora no

obtuvo ningún retweet su valor será de 0. A partir de los datos de entrenamiento evalúan un conjunto de tweets y predicen cuáles tendrán mayor éxito (más RT) en un periodo de una hora, obteniendo la correlación del interés de un grupo de usuarios y los tweets publicados. Los autores mencionan que su metodología puede ser mejorada debido a su flexibilidad en el manejo de los datos debido a que permite incorporar nuevos criterios que aumenten la correlación de los tweets y el interés de los usuarios.

Otra contribución importante es la realizada por Ma et al. [35], quienes proponen un método que pretende predecir qué tópico puede convertirse en popular en un futuro cercano, formulando el problema desde un enfoque de clasificación utilizando Naive Bayes, KNN, Árboles de Decisión, Máquinas de Soporte Vectorial y Regresión Logística. En su investigación las características no se limitan al HT, también consideran las propiedades de los mensajes relacionados al tópico, como es el caso de los enlaces, sentimiento del tweet, la acción (RT o mención) entre otros. En total examinan cinco características del HT a predecir (número de HT incluidos en el tweet, categoría del HT, número de veces en que se mencionó el HT, etc), once características relacionadas a los tweets que integran al HT (polaridad del tweet, número de enlaces incluidos en el tweet, número de usuarios mencionados, etc.) y a los usuarios que publicaron (número de seguidores por usuario, geolocalización, total de veces que utilizaron el HT, etc). Además, definen cinco rangos para explicar el crecimiento de un HT, que son: no popular, marginalmente popular, popular, muy popular y extremadamente popular. Con estas descripciones los autores clasifican a los HT en el tiempo. Concluyen que su trabajo es uno de los primeros en modelar y cuantificar la popularidad debido a que predicen el número de usuarios que pueden unirse a la conversación por medio del HT. A diferencia de Ma et al., la investigación realizada por nosotros utiliza un algoritmo incremental debido a la cantidad de instancias obtenidas, de manera que no es necesario generar el árbol en cada solicitud de nuevos tweets, únicamente el modelo se reconstruye si hay cambios significativos, favoreciendo el rendimiento de la plataforma y adaptándose a posibles nuevos atributos integrados por la misma RS.

Adicionalmente a los trabajos anteriores se han explorado métodos propuestos por diversos autores, para así posicionar la propuesta posteriormente planteada. Entre los primeros estudios del análisis de las tendencias y su propagación se encuentra el framework denominado “Conversational Vibrancy” propuesto por Lin et al. [36], dónde explican el crecimiento, duración y el contexto de 256 HT durante el debate de los Estados Unidos efectuado en el 2012. De acuerdo a su duración clasifican a los HT como “winners” y “also-rans”, la primera categoría se caracteriza por emerger rápidamente, y la segunda por ser mencionado durante periodos más largos. Su investigación es un primer acercamiento a explicar el ciclo de vida de un HT y proponen cuatro elementos generales: i) actualidad (topicality), ii) interactividad, iii) diversidad y iv) prominencia.

La actualidad (i) describe la relevancia del momento y contexto: los tweets etiquetados con los HT están más propensos a crecer y persistir de manera que son incluidos en la conversación. La interactividad (ii) es cuando los usuarios responden a un tweet o mención; están predispuestos a crecer si los usuarios interactúan con otros usuarios e incluyen el HT en la conversación. La diversidad (iii) captura la actividad de los tweets relacionados al HT, es más sencillo el crecimiento cuando hay variedad en el contenido de los tweets. Por último, la prominencia (iv), mide la audiencia expuesta en el HT, aquellos usuarios mencionados con mayor número de seguidores es más probable que ayuden en el crecimiento comparado con aquellos que tienen pocos seguidores. Desde un enfoque estadístico analizan el incremento en popularidad y persistencia de los HT considerando los cuatro factores antes mencionados, y exhiben el nivel de crecimiento basado en el RT, interactividad, diversidad de contenido y número de participantes.

Altshuler et al [37] presentan un modelo analítico basado en modelos de difusión de información y enfocado en el análisis de las interacciones sociales pasadas para predecir las futuras. Su estudio presenta una evolución de la difusión de las tendencias y patrones observados durante su investigación. Su pregunta de

investigación es: ¿cuál es la probabilidad de que los mensajes produzcan una difusión viral y una difusión generalizada?, los autores mencionan que el factor clave para tener éxito es la optimización de los mensajes, en el sentido de determinar cuál es el mejor camino para convertirse en tendencia. Su principal aporte es un modelo fundamentado en procesos de difusión y redes libres de escala, de manera que consideren muchos caminos y se puedan vincular de manera global. Aunque no presentan un caso de estudio de tweets extraídos, sí presentan teoremas que fundamentan su modelo analítico.

Zhang et al. [38] investigan dos cuestiones básicas en la predicción de tendencias: por un lado cuáles son los factores importantes, y por el otro, cuáles pueden ser los modelos adecuados para esta tarea. Los autores se enfocan en diferentes factores de contexto y contenido como es el caso del tweet, la topología de red, el comportamiento del usuario, etc. Uno de los modelos que les dio mejor resultado fue Random Forest para el análisis de las características importantes. Concluyen que el contenido y los factores de contexto ayudan a la predicción de tendencias, sin embargo, el comportamiento del usuario en la tendencia y su actividad en ella es más importante. En cuanto al estudio de sus modelos de predicción mencionan que los no lineales son significativamente mejores que los lineales. De igual manera que otros autores no intenta posicionar un HT, se enfocan principalmente en el estudio de patrones y determinación de factores que posibiliten la viralidad.

Con base en lo expuesto en esta sección, podemos afirmar que existen importantes vacíos en el trabajo llevado a cabo hasta el momento, sin que esto implique que éste sea irrelevante. Desde la perspectiva que nosotros abordamos la problemática del análisis y explotación de Twitter como plataforma de RS, uno de los temas de mayor interés para efectos de nuestro trabajo es la predicción, siendo ésta todo un reto por sí misma debido a que los datos que se recopilan están constituidos por muchas variables de diferente tipo, y la cantidad de datos a adquirir de manera constante requieren de un pre-procesamiento para hacer factible su uso.

Diversos autores se han interesado por el estudio de la predicción en Twitter, sin embargo se han abordado principalmente desde un enfoque de ventas, análisis de situaciones de desastres naturales, y el éxito en campañas políticas; a su vez se han propuesto diferentes modelos para deducir cuál es la actividad, las características de las tendencias y las razones de su viralidad. En nuestro caso, nos enfocamos en el estudio del HT y las características que determinan que lo hizo popular o desapercibido.

### *2.4. Clasificación incremental.*

Recientemente los datos de transmisión, o en inglés conocidos como “stream data” ,se han convertido en un tema de investigación de gran importancia, este tipo de datos se ordenan en secuencia de instancias a mayor velocidad. Las RS son un claro ejemplo de *stream data* que involucran millones de datos con alta dimensionalidad y de constante cambio, que hace que la clasificación tradicional se vuelva imprecisa y computacionalmente costosa. Los árboles de decisión (*decision trees*) son considerados como modelos populares y valiosos para la clasificación y el proceso de predicción. En su mayoría son fáciles de usar y pueden manejar adecuadamente la incertidumbre, además proporcionan recomendaciones para la toma de decisiones y proveen un esquema para cuantificar el costo de un resultado y la probabilidad de que esto suceda. Entre sus inconvenientes se encuentran que cuando el número de alternativas es grande los caminos obtenidos pueden ser complicados de comprender. Hay modelos populares de DT como son ID3 (Interactive Dichotomiser 3) [39], C4.5 [40] es una extensión de ID3 y C5.0 una mejora del anterior, entre otros, sin embargo no están preparados para *stream data*.

Debido a esta necesidad los autores Domingos y Hulten [41] propusieron VFDT por sus siglas en inglés “Very Fast Decision Tree”, es un algoritmo de decisión popular para minería de flujo de datos o *data stream mining.* El proceso de construcción del algoritmo se basa en el principio de la cota de Hoeffding [42] el cuál decide si dividir los nodos de acuerdo a la estadística de los datos de sus hojas. Aunque existen

variantes de VFDT como M-VFDT o CVFDT nos enfocamos a presentar únicamente investigaciones que utilizan VFDT para resolver problemas de flujo de datos constante.

Zhang et al. [43] describen el diseño de un sistema de minería de datos médicos que realiza predicción en tiempo real. En su diseño cada nodo se puede considerar como una etiqueta de clase e indica una situación médica o enfermedad, cuando existe un nuevo registro el algoritmo de predicción usa la entrada para encontrar registros similares y así recibir un posible tratamiento de la enfermedad.

Mencionan que la mayoría del software de este tipo únicamente pueden analizar conjuntos de datos finitos y estructurados, y es en ese caso que los árboles de decisión tradicionales puede ser una manera de adecuada de resolver el problema; sin embargo, en la medicina los flujos de datos son rápidos por lo que decidieron utilizar un árbol más rápido sin perder la exactitud, y es por eso la selección de VFDT fue la adecuada. Concluyen que la aportación principal es su utilidad en casos de emergencia o en misiones de rescate debido a que efectúa la predicción en tiempo real y con datos en línea.

Minegishi et al. [44] presentan el uso de VFDT en un problema que considera datos reales y constantes como las transacciones bancarias para la detección oportuna de fraudes en tarjetas de crédito. La ventaja que tuvieron al utilizar VFDT es que los ejemplos que obtenían durante el flujo de datos no se acumulaban en memoria, es conforme llegaban los nuevos ejemplos el árbol crecía gradualmente acumulando la información estadística en el nodo anterior y midiendo si los nuevos nodos cumplían con los criterios estadísticos de la cota de Hoeffding. Las transacciones que los autores analizaron eran complejos y con tiempos cambiantes en la recepción de los datos, además que debían considerar casos reales como: i) aproximadamente un millón de datos por día, ii) cada transacción lleva menos de un segundo, iii) aproximadamente llegan cien transacciones por segundo en horas de mayor transferencia, iv) las transacciones llegan las 24 horas del día. Las características

mencionadas pueden ser comparadas con el flujo de información que se puede tener en Twitter. En su propuesta comparan los algoritmos C4.5 y VFDT. Para el primero fue necesario recolectar todos los datos y posteriormente generar el árbol logrando una precisión de 92.980%, mientras que VFDT generaba el árbol conforme llegaban los datos y el cuál tuvo 95.459% de precisión. Como conclusión mencionan que para cierto tipo de datos C4.5 es mejor aunque más lento y necesita tener todos los datos recolectados y VFDT por sus características puede generar el árbol conforme llega la información y suele ser más rápido que el primer algoritmo.

Limón et al. [45] describen un nuevo sistema de minería de datos distribuidos basados en el paradigma de sistemas multi-agentes y artefactos. En él evalúan diferentes estrategias centralizadas y de rondas como es el caso de VFDT. En las estrategias por rondas se enfocan a aprender un modelo inicial con todos los ejemplos que tienen disponibles en un nodo y posteriormente el modelo se mueve al siguiente nodo para actualizar de acuerdo a cada ronda, el proceso es continuo hasta que se terminan las rondas asignadas. Los autores utilizaron 18 conjuntos de datos con variaciones en el número de atributos, instancias y clases, y así obtener resultados comparables entre los algoritmos de su selección.

Los autores mencionan que VFDT se comporta sutilmente mejor que J48 para las clases binarias, como es en nuestro caso en los HT, si fue o no tendencia, además que VFDT es más rápido que J48. Consideran que la raíz del problema de VFDT es que aunque es incremental no se puede mover una vez que el nodo ha sido creado, por lo cuál en un futuro desean evaluar una variante llamada CVFDT.

Por último, otra investigación relevante es la de Dinata et al. [46] quienes proponen un *framework* que utilizan el clasificador incremental VFDT para un problema de flujo constante como es la información de sensores en una casa inteligente. En la primera etapa aplican métodos de pre-procesamiento para reducir el ruido en los datos y valores faltantes, y posteriormente utilizaron VFDT para mejorar la eficiencia en la clasificación reconstruyendo el árbol únicamente si las ramas tuvieron cambios

significativos en su estadística. De esta manera lograron reducir el tiempo de entrenamiento obtenido de los sensores y generando una predicción con mayor precisión. Los autores concluyen que los datos de flujo constante tienen diversos problemas como son la alta dimensionalidad y el tamaño de la base de datos, por lo que la integración de VFDT a su propuesta apoyó en su proceso de clasificación sin que la precisión se viera afectada.

Las investigaciones mencionadas son un ejemplo de que el algoritmo incremental seleccionado para nuestra propuesta trabaja adecuadamente con *stream data* o datos en línea, que es el caso de los tweets obtenidos. Como será explicado en los siguientes capítulos, nuestra propuesta toma como inspiración algunos de los elementos de las investigaciones de otros autores, sin embargo nuestro enfoque es sustancialmente diferente, y el objetivo al que pretendemos llegar, es el determinar dinámicamente los factores que hacen que un HT establecido sea exitoso, siendo diferentes de los demás trabajos analizados.

# CAPÍTULO III: Metodología

Tras analizar las propuestas actualmente existentes en este contexto de trabajo, tanto en términos académicos, como de productos comerciales, hemos llegado a la conclusión de que nuestra investigación es significativamente diferente a las demás. La razón principal es que nosotros proponemos un método de predicción de tendencias y lo abordamos como un problema de clasificación. Una particularidad importante que hay que mencionar es que, dada la cantidad de información que se maneja, y que ésta se encuentra en constante cambio, hemos optado por utilizar una aproximación basada en un algoritmo incremental sobre una plataforma de sistemas multiagentes (SMA), pues de esta manera se consideró que se podrían lograr mejores resultados, como se describirá a detalle más adelante.

El argumento para hacer uso de un algoritmo incremental se justifica en el hecho de que la cantidad de datos que se analizan es muy grande, y está en constante crecimiento y su forma es considerablemente cambiante, motivo por el cual se vuelve impracticable llevar a cabo un nuevo análisis de los datos cada vez que éstos se actualizan.

En lo que concierne a los agentes, se consideraron un medio eficiente a través del cual se pueden organizar todas las diferentes acciones que se requieren llevar a cabo para lograr la obtención de resultados. Dichos agentes se encargan de realizar las fases de extracción, pre-procesamiento, clasificación, comparación de los modelos generados, creación de reglas de asociación y publicación automatizada de los tweets.

Los HT, como ya se mencionó previamente de manera somera en este mismo documento, son la pieza clave de la investigación, a partir de la cual determinamos los elementos principales que dan como resultado las tendencias en Twitter. Si bien, es conocido que para crear un TT es fundamental tener un vasto número de RT realizados por un grupo importante y variado de usuarios en periodos cortos de

tiempo, también existen otros elementos que hay que tomar en consideración. El problema es que éstos no son del todo evidentes, razón por la que los denominamos factores ocultos. Determinar qué tipo de mensaje, qué tópico va a crear interés, y que por lo tanto, se compartirá en Twitter, es muy importante en muchos campos de investigación. Por ejemplo:

- ¿Quiénes prefieren dar RT a un tweet que incluye videos?
- ¿Cómo generar una campaña de publicidad dirigida en determinado tiempo?
- ¿Se distribuirá más un tweet sobre un tema político que uno sobre un tema deportivo?

El incremento constante de la cantidad de datos disponibles en Twitter da como resultado que su análisis se vuelva una tarea extremadamente difícil, razón por la cual es imperativo hacer uso de técnicas innovadoras de análisis de datos, como es la Minería de Datos (MD). Si se usa adecuadamente, es posible llegar a identificar patrones sobre la forma en que se comparte la información, las características de los usuarios y los temas más populares, entre otros. Estos algoritmos, al ser usados en conjunto con otros de Aprendizaje Automático, permiten "*aprender*" posibles relaciones entre diversos factores. El hecho de que Twitter permita la recuperación de tweets del pasado reciente facilita el proceso de análisis, pues con base en ese análisis del pasado se pueden obtener proyecciones en el futuro, lo que es altamente útil para nuestro trabajo.

Con base en estos elementos fundamentales, nosotros proponemos una metodología a través de la cual es posible llevar a cabo la recolección y el análisis de los datos básicos necesarios para el presente trabajo, cuyo objetivo es determinar los factores que dan como resultado que un tweet determinado pueda llegar a tener un mayor grado de viralidad. De acuerdo con Guerini [47], actualmente la viralidad es simplemente una forma acelerada y mejorada de el mensaje de boca en boca, sin embargo Hansen [48] argumenta que es algo completamente diferente, en dónde el "virus" o mensaje propagado está directamente relacionado con el número de usuarios que atrae. Para Twitter, se refiere a aquellas publicaciones que, al difundirse

obtienen una gran cantidad de RT o comentarios dentro del mensaje original en un periodo corto.

De esta manera la idea es contar con un sistema de generación de recomendaciones, que sea capaz de, tomando como base el análisis de la situación pasada y actual de Twitter, el tema que se pretenda abordar en el tweet que se piensa emitir, y otros factores básicos, determinar las posibles reglas que se podrían aplicar para efecto de lograr que el citado tweet logre un mayor alcance en términos de los usuarios a los que les cause algún interés, o lo que es lo mismo, que tienda a propagarse “viralmente".

A partir del análisis que llevamos a cabo de los tweets, nuestra propuesta tendría la capacidad de identificar las características que pueden facilitar que un tweet sobre un tema determinado sea más popular, generando recomendaciones sobre las características que deberán satisfacer los siguientes tweets, a ser publicados de manera semi-automática, dando como resultado, a final de cuentas, la generación de recomendaciones.

El método de recomendación que nosotros proponemos puede ser parametrizado para adaptarse a nuevos escenarios o atributos, como el análisis de tweets en inglés y considerar más categorías, sin embargo las características definidas para el estudio que nos atañe son las siguientes:

1. Los HT analizados únicamente pertenecen a las categorías “Deportes”, “Espectáculos” y “Política”, según el modelo de clasificación de [23];
2. Los HTs estudiados se geolocalizan específicamente en México;

Como se explicará a detalle en las siguientes secciones del presente capítulo, los tweets extraídos se someten a un proceso de pre-procesamiento, a través del cual se elimina información poco relevante para la investigación, como es el caso de los *emoji, palabras vagas (stop words), menciones, e*ntre otros elementos. Después se

hace una selección de atributos y se realiza el aprendizaje mediante un árbol de decisión incremental, y finalmente, obtener reglas de asociación que el usuario puede interpretar como recomendaciones.

A continuación detallamos la estructura del método y las etapas de nuestra propuesta.

### *3.1. Definición inicial de los parámetros*

En la etapa inicial de este proyecto, durante un periodo correspondiente a aproximadamente seis meses, se realizó un estudio empírico preliminar de los HT mencionados en México, y específicamente de aquellos que llegaron a ser TT. Esto nos permitió explorar, de forma incipiente, la manera en que se lleva a cabo el flujo de la información, así como la forma en que se producen las interacciones de los usuarios. En esta primera etapa observamos el comportamiento de cincuenta HT, a partir de los cuales logramos encontrar patrones de interés por parte de los usuarios de Twitter. Para ello diseñamos un algoritmo para recopilar los HT al azar, como regla general se obtuvieron en un periodo de cinco días y ejecutando el proceso descrito a continuación en tres momentos diferentes. Éstos corresponden al día (de 9 a 11 hrs), tarde (de 15 a 17 hrs) y noche (21 a 23 hrs), el algoritmo se ejecutó de manera aleatoria para evitar la extracción de los HT siempre el mismo día a la misma hora. Las fases del algoritmo se dividen en dos y se describen a continuación.

Primera fase: extracción de tendencias tendencias:

1. Seleccionamos de manera aleatoria un número (*n*) del 1 al 5
2. Generamos *n* veces un nuevo número aleatorio (*pos*) del 1 al 10
3. Conseguimos de la lista de tendencias de Twitter la posición adquirida de manera aleatoria (*pos*) en el paso dos
4. La tendencia correspondiente a la posición se guarda en una colección de datos

5. Se repite hasta cumplir con el total de *n* de la fase uno.

Segunda fase: derivado de la recopilación de tendencias, los pasos para adquirir los HT menos populares son los siguientes:

1. Seleccionamos de manera aleatoria un número (*n*) del 1 al 5
2. Generamos *n* veces un nuevo número aleatorio (*pos*) del 11 al 50
3. Conseguimos de la lista de HT de Twitter la posición adquirida de manera aleatoria (*pos*) en el paso dos
4. El HT correspondiente a la posición se guarda en la misma colección de datos que las tendencias
5. Se repite hasta cumplir con el total de *n* de la fase uno.

Los pasos anteriores se repitieron hasta tener el total de cincuenta HT para iniciar el estudio de los tópicos y encontrar pautas que describieran los intereses de los usuarios. Posteriormente, clasificamos los HT de manera manual considerando la investigación y propuesta de Romero et al. [23].

Los tópicos seleccionados de manera aleatoria para su estudio se enfocaron principalmente a temas relacionados con los deportes, la política y el entretenimiento, y quedaron al final temas relacionados a la ciencia, la comida y la literatura. De manera anecdótica podemos señalar que en países como Inglaterra o Canadá es más probable encontrar el tópico tecnología como tendencia. Se hallaron tres casos atípicos de comportamiento que son:

a) Durante el estudio regularmente se observó que los tópicos de la categoría “Literatura” sobresalían, volviendo rápidamente a un comportamiento más regular de bajo perfil, y de poca aparición dentro de Twitter en el contexto de análisis. Después de analizar los tweets de la categoría “Literatura”, se observó que estas apariciones se daban de manera meramente incidental debido a efemérides relacionadas con temas de literatura, y a un hecho muy particular, y fue que en

esos momentos se celebraba la Feria Internacional del Libro en la ciudad de Guadalajara. Fuera de ello, este no es un tema recurrente o que se mencione en Twitter en México.

b) La categoría de “Comida” también se menciona en México, pero resultó estar correlacionado con la categoría de “Espectáculos”, esto debido al programa de televisión llamado Master Chef, que se transmitía durante al menos parte de las fechas del estudio. Logramos percatarnos de que el tópico principal es “Espectáculos”, y posteriormente el tema “Comida” deriva a partir del primero.

c) El tema de la tecnología se posiciona regularmente como tendencia en México, también de manera asociada a eventos específicos cuando, por ejemplo, en Estados Unidos se está celebrando un evento tecnológico, como lo es el “Consumer Electronics Show” (CES), o la “Game Developers Conference”, por ejemplo.

Con base en los resultados obtenidos y descartando los casos destacados por un tema tercero, seleccionamos las categorías más populares que describen los patrones de publicación de los usuarios en México. En caso de que el método sea utilizado para otro país o idioma se deberá hacer inicialmente un estudio de los tópicos populares del país a analizar.

### *3.2. Extracción*

La primera etapa realmente efectiva dentro del proceso de análisis de Twitter corresponde a la extracción de los tweets, tarea no del todo trivial. Twitter ofrece tres medios a través de los cuales los usuarios programadores pueden acceder a los datos que contiene. Se dan bajo la forma de algo conocido como Interfaz de Programación de Aplicaciones (API - Application Programming Interface). Estos medios son: *REST API*, *Streaming API* y *Search API*. Cada uno de ellos tiene características particulares. Dependiendo del método que se utilice, es necesario autenticarse para poder obtener los datos que pueden ser en formato XML

(eXtensible Mark-up Language), JSON (JavaScript Object Notation), RSS (Really Simple Syndication) o Atom.

La *REST* (REpresentational State Transfer) [49] *API* corresponde a un enfoque de desarrollo y uso de servicios Web definido por Roy Fielding. Este servicio ofrece a los programadores el acceso a solicitudes autenticadas por Twitter. Todas las operaciones de extracción se realizan mediante GET y POST, que son métodos propios de HTTP. En este caso se puede seleccionar cualquiera de los cuatro formatos antes mencionados.

La segunda alternativa es la *Streaming API*, la cual proporciona subconjuntos de tweets en un modo de operación cercano a tiempo real. Esto se hace a través del establecimiento de una conexión permanente en cada petición, y el flujo de datos se alimenta cada vez que otros usuarios publican tweets que satisfacen los criterios determinados por el programador en el momento de establecer la conexión. A diferencia de la *REST API*, esta alternativa recibe el flujo de tweets exclusivamente en formato JSON. Permite obtener datos aleatorios, filtrados por palabras clave, o exclusivamente los tweets de un usuario dado. El formato JSON el cual describe los datos con una sintaxis dedicada que se usa para identificar y gestionar los datos, además de ser más ligero que un archivo XML [50].

Por último, la *Search API* provee tweets de acuerdo a una cadena de búsqueda indicada a través de una petición Web. No requiere una autenticación y los tweets pueden obtenerse en formato JSON o Atom. Su desventaja es la limitación de los tweets y la información proporcionada ya que, a diferencia de las dos anteriores, no ofrece un perfil completo del usuario que ha realizado la publicación y que permiten obtener información más completa del usuario que emite los tweets. Para efectos de la presente investigación, se decidió hacer uso de la *REST API*, a través de la cual se obtienen tres instancias diferentes: a) información completa del tweet, b) información completa del usuario que lo publicó, y c) el número de seguidores y el total de sus respectivos seguidores.

Dado este hecho fue menester autenticarse mediante el proceso denominado OAuth [51]. OAuth es un protocolo abierto que permite autorización segura de una API utilizando métodos simples para aplicaciones Web, escritorio y móviles. Es una de las formas actualmente más populares que permite al usuario final otorgar accesos del uso de los métodos, en este caso, de Twitter. Actualmente, existen diversas implementaciones de OAuth en diferentes lenguajes de programación, lo que facilita su integración a las aplicaciones.

Continuando con el flujo de autorización en el portal de Twitter es necesario crear un espacio de trabajo conocido dentro del contexto de la RS como aplicación, la cuál le asignamos el nombre de “TweetPop, y posteriormente se nos concedieron las llaves de acceso para el uso de su API. Las llaves de acceso se componen de un “Consumer Key” y un “Consumer Secret”, ambas son cadenas únicas encriptadas generadas por Twitter. La solicitud de autenticación consiste en utilizar dichas claves en cada petición y consumo de los diferentes métodos de la API.

Una vez adquirida la autorización de Twitter hicimos una primera selección de los atributos del tweet y definimos los HT que fueron de nuestro interés analizar. Para finalizar, diseñamos un algoritmo que extrajo una muestra aleatoria de los tweets que contenían en su mensaje el HT a observar. El servicio de Twitter limita la extracción a 200 tweets por solicitud, por lo que es necesario moverse entre las páginas que usa Twitter para agrupar los mensajes. Se describen a continuación el algoritmo de recolección de los tweets:

1. Utilizamos el servicio de Twitter para la búsqueda de tweets
2. Asignamos como parámetro de búsqueda el HT a observar
3. Obtenemos un valor aleatorio de 1 a 10 denominado *n*
4. Desde uno hasta n
    1. Nos posicionamos en la página *n* de la respuesta proporcionada por Twitter
    2. Seleccionamos de manera aleatoria hasta 150 tweets

3. Los tweets seleccionados se guardan en la colección de datos

5. Finalizamos la extracción

Los tweets recolectados están compuestos por una serie de atributos que se agrupan de acuerdo a lo que describen, también denominadas entidades. Entre las entidades que brinda Twitter se encuentra la información general del usuario, que elementos componen al tweet, datos de geolocalización en caso de que el usuario la tenga activada y la descripción de la cuenta que publicó. Para nuestro estudio no son necesarios todas las entidades, y únicamente utilizamos aquellas que aportan datos relevantes a la investigación.

En este punto cabe mencionar que la información general del usuario está constituida por treinta y siete atributos diferentes, y la información concerniente al tweet está conformada por tres conjuntos de datos independientes. Primero, las entidades que describen los elementos que componen el tweet, como es el caso de las URLs y el HT; en segundo lugar, las entidades extendidas que son atributos descriptivos de los elementos multimedia que se encuentran en el tweet, como el caso de imágenes, GIF, videos, etc; normalmente cada uno está conformado por quince atributos. Por último, se encuentra una entidad denominada “places”, la cual contiene los lugares con la descripción de las coordenadas GPS, en caso de que el usuario tenga activada su geolocalización.

En total, se obtuvieron noventa y dos atributos. Entre ellos, fue necesario considerar sólo los que aportaban efectivamente información relevante para nuestro estudio, y descartar algunos de ellos, como es el caso de “profile_url”, que contiene la dirección electrónica del usuario asignada por Twitter, “profile_background”, que contiene el nombre de la imagen seleccionada por el usuario para su perfil. En lo concerniente a los demás atributos disponibles, su utilidad no es necesariamente evidente. En ese caso es necesario preparar los datos que puede generar un conjunto de datos más pequeños y menos atributos pero con mejor calidad en la información, por lo que es necesario utilizar técnicas para pre-procesar los datos.

### *3.3. Pre-procesamiento*

El pre-procesamiento es la etapa a través de la cual se busca asegurar la calidad de los datos que hay que procesar para evitar que haya información duplicada que afecte a la predicción y, además ahorre espacio y tiempo en el análisis. Engloba a todas aquellas técnicas de análisis de datos que permite enriquecer la calidad de los conjuntos de datos de manera que la etapa de extracción tenga una mejoría en los datos, y así obtener información relevante para el estudio. Además, se agiliza la consulta y la clasificación de los datos. Como parte del pre-procesamiento se aplicaron algunas reglas de unificación y reducción de la dimensión, como eliminar los atributos irrelevantes o redundantes. También se busca reducir la numerosidad, que no es más que la representación de datos grandes a través de valores más pequeños, como agregar un 1 si el usuario incluye algún elemento de multimedia a su contenido.

Para efecto de poder identificar y extraer los HT, por razones evidentes, pues a final de cuentas todo se reduce a estructuras sintácticas, se decidió hacer uso de expresiones regulares, pues a través de ellas es posible describir de manera detallada las características que permiten identificar a este tipo de componentes, cruciales para nuestro trabajo.

Otra técnica utilizada para el pre-procesamiento fue la de Latent Direchlet Allocation (LDA) [52, 53], que no es más que un procedimiento a través del cual las palabras que forman parte de un tweet se agrupan según la categoría a la que pertenecen. En este caso en particular, se aplicó para la limpieza de los tweets. Es una etapa crucial en el análisis porque no se deben perder datos importantes, sino más bien eliminar aquello que se considera poco útil o *basura* en los datos. Se utilizó la técnica de limpieza de “stop words” (palabras vagas o vacías en español), que son palabras que no aportan un significado al mensaje y que deben eliminarse para quedarse exclusivamente con las que contribuyan al contenido del mensaje. Uno de los

problemas principales con los que nos encontramos fue la imposibilidad de localizar un diccionario, o corpus, de palabras vagas en español. Para ello se utilizó un diccionario base y posteriormente se fueron incluyendo nuevas palabras que consideramos innecesarias para el análisis. Las menciones y URL se extrajeron del mensaje para después almacenarse en la base de datos de manera independiente, para posteriormente asignarles el atributo **`siteType`** antes mencionado.

Una vez que la información está lista para el análisis, se lleva a cabo un proceso de clasificación mediante un algoritmo incremental, el cual se detalla en el siguiente punto.

Esto es importante porque la selección de atributos que hay que tomar en consideración es crucial, y se debe eliminar aquellos que aporten muy poco y conservar los que efectivamente puedan ser considerados como determinantes. De no hacer esto, se corre el riesgo de que los atributos poco relevantes introduzcan *ruido* en el proceso de análisis y pueda dar lugar a la generación de conclusiones erróneas.

### *3.4. Selección de atributos*

Las técnicas de selección atributos reducen el costo computacional asociado al aprendizaje y que pueden confundir a algunos clasificadores, además que aumentan la precisión eliminando aquellos atributos que pueden ser irrelevantes para el aprendizaje. Uno de los objetivos principales de la selección de atributos es encontrar el subconjunto mínimo de atributos tal que no se afecte significativamente la precisión de la clasificación.

Para la selección de atributos utilizamos una herramienta de aprendizaje automático y minería de datos llamada Weka, siendo una de las más populares por su facilidad de uso, su amplia gama de algoritmos de aprendizaje automático y la robustez que ofrece para el análisis. Además está conformada por colecciones de herramientas de

visualización, técnicas de procesamiento de datos y modelado. Para la medición de la calidad de los atributos mediante Weka aplicamos la técnica denominada Ganancia de Información (GI) [54]. Ésta es una medida de cuánto ayuda el conocer el valor de una variable aleatoria *X* para conocer el valor de *Y,* donde *X* es un atributo de un ejemplo dado, y *Y* es la clase a la que pertenece el ejemplo. Se busca una ganancia alta que implica que el atributo *X* permite reducir la incertidumbre de la clasificación.

Usando Weka aplicamos la técnica GI a una muestra preliminar para la selección de atributos que apoye en nuestro estudio de los datos. Como resultado tenemos los atributos mostrados en la tabla 1. La primer columna representa el puntaje de clasificación y la segunda columna el nombre del atributo del conjunto de datos seleccionados.

| Puntaje de clasificación | Nombre del atributo |
|---|---|
| 1.381901 | totalRT |
| 1.236707 | typeContent |
| 0.2882709 | dayInterval |
| 0.1945599 | day |
| 0.0924448 | seasonal |
| 0.0807265 | followersCount |
| 0.051562 | resultType |
| 0.043202 | additionalContent |
| 0.0015093 | category |

*Tabla 1. Puntaje de clasificación de los atributos seleccionados*

Weka lo que hace es a) seleccionar del total de variables que entregan mayor cantidad de información, b) se seleccionan el resto de las variables disponibles que de igual manera transmiten la mayor cantidad de información nueva y c) se repiten

los pasos a y b hasta que la ganancia de información ya no pueda justificar la pérdida de los datos. El conjunto obtenido tiene las siguientes características:

- Provee de la mayor cantidad de información dado los datos disponibles,
- Las variables seleccionadas no son colineales, es decir que son independientes y la aportación de información es independiente
- Se tiene en cuenta las interacciones entre variables
- Se obtienen variables que por el experto son consideradas como importantes, sin embargo, el objetivo es que sean seleccionadas mediante una técnica que respalde su elección.

Así, en una primera instancia descartamos los atributos que a nuestro parecer son de poca utilidad (ejemplo: imagen de fondo) y posteriormente se hizo una selección utilizando Weka. Una vez obtenidos los resultados aplicando GI los atributos que resultaron ser los más relevantes para el nuestro estudio, se presentan en la tabla 2,

| Atributo | Descripción |
|---|---|
| categoría | Es el asignado al HT evaluado |
| tipo de tweet | Es el tipo de tweet que puede ser popular, mixto o reciente |
| intervalo del día | El intervalo de tiempo en que fue publicado. |
| día de publicación | El día de la semana en que fue publicado |
| contenido adicional | Toma el valor de 1 si tiene contenido adicional y 0 en caso contrario |
| tipo de contenido adicional | Puede ser video, enlace, imagen estática o todos los anteriores |
| total de RTs | Es el número de RTs que obtuvo cada tweet |
| total de seguidores | Es el número de usuarios que siguieron al TT e interactuaron con él |

*Tabla 2. Atributos finales*

Tras esta primera selección, se agregaron nuevos atributos del estudio de los patrones analizados a lo largo del proceso de investigación. Éstos son denominados atributos derivados debido a que Twitter no los proporciona directamente, más bien

se deben calcular y almacenar posteriormente. Ejemplos de este tipo de atributos es el “tamaño de la palabra” (**`wordLenght`**), un valor numérico que corresponde al total de palabras concatenadas que forman el HT. En el caso de que el evento se repita cada cierto tiempo por ejemplo Navidad, día de la Bandera, etc el atributo es (**`seasonal`**) Además se incluye “tipo del sitio” (**`siteType`**), que clasifica el enlace, en caso de que en el tweet se haya incluido alguno. Los valores posibles pueden ser: *RS*, *portalA*, *portalB*, *imagen* y *streaming*, que se refieren a sitios de distribución digital de contenido. El valor *RS* se refiere a que el contenido adicional corresponde a un enlace de una Red Social como es el caso de una imagen en Instagram, un video en Youtube, una publicación en Facebook, entre otros. Los valores para *portalA* y *portalB,* se refieren a los enlaces provenientes de publicaciones en blogs o sitios de noticia, si corresponden a un portal de información seria es *portalA*, en caso de ser un enlace a un sitio de “chismes”, parodias, burlas, etc, entonces se clasifican como *portalB.* El valor *imagen* se asigna cuando en su el tweet viene insertada, y por último *streaming* son aquellos denominados de transmisión continua.

La clasificación del tipo del sitio (**`siteType`**) se realizó de manera semi automática considerando las meta-etiquetas que incluía en su contenido Web, para ello proponemos un algoritmo que extrae las palabras clave de la meta-etiqueta de cada sitio y comparando con un diccionario de datos que hicimos de manera manual se evalúa cuántas de las palabras corresponden a cada categoría antes mencionada Se realiza un cálculo y la categoría con mayor puntaje es la asignada a la etiqueta del tweet. Describimos a mayor detalle el algoritmo considerando que el diccionario de datos ya ha sido creado.

1. Extraemos el tweet
2. Si el tweet tiene contenido adicional
3. Obtenemos el dominio del contenido
4. Si el dominio no se encuentra en el conjunto de datos clasificados, entonces
    1. Extraemos la meta-etiqueta de tipo *“keywords”*
    2. Comparamos cada meta-etiqueta con el diccionario de datos

3. La etiqueta que tiene mayor número de apariciones es asignada al tweet y al dominio

5. En caso contrario asignamos la etiqueta que se encuentra previamente asignado en el conjunto de datos al tweet en el atributo **`siteType.`**

Por último, la “clasificación del usuario” (**`uInfluence`**), es asignada a cada usuario que participó en el HT mediante un nombre descriptivo o también denominado *etiqueta lingüística*.

Las etiquetas lingüísticas son variables cuyos valores pueden ser representados mediante conjuntos difusos, todos los valores lingüísticos forman un conjunto de términos o etiquetas.

En nuestro caso, el universo es el número de seguidores en Twitter y la etiqueta asignada corresponde al conjunto difuso dependiendo del número de seguidores que tiene el usuario que publica el tweet.

Una de las ventajas de usar etiquetas lingüisticas se conoce como granulación, que es la forma de comprimir la información. El objetivo es caracterizar fenómenos que no están bien definidos, en este caso de acuerdo al rango de seguidores, ya que no es posible saber a ciencia cierta el nivel de difusión que puede tener un usuario. Es por ello, que trasladamos ese poder de difusión a un valor numérico que, a su vez, tiene asignada una etiqueta que permite clasificarlo de una manera más simple [55].

A mayor número de seguidores existe un mayor número de posibilidades de que otros usuarios lean un tweet o RT, es por ello la importancia de determinar su etiqueta. Los seis valores lingüísticos que definimos se muestran en la Tabla 3.

| Etiqueta | Rango de seguidores |
|---|---|
| Desconocido | 0 a 1,000 |

| | |
|---|---|
| Ordinario | 1,001 a 10,000 |
| Destacado I | 10,001 a 100,000 |
| Destacado II | 100,001 a 1,000,000 |
| Destacado III | 1,000,001 a 10,000,000 |
| Famoso | 10,000,000 y más |

*Tabla 3. Etiquetas lingüísticas*

Como se puede apreciar claramente en esta tabla, el valor lingüístico *"Desconocido"* representa a los usuarios que por lo regular acaban de crear una cuenta o que tienen poca actividad y casi nula interacción con otros usuarios. Los usuarios «Ordinarios» son aquellos que comienzan a ganar popularidad y, en consecuencia, comienzan a ganar seguidores interesados en sus publicaciones y RT. Los usuarios *"Destacados"* se dividen en tres niveles, en función de su número de seguidores; son cuentas de usuarios más activos que realizan diversas publicaciones durante el día, de interés para otras personas, por lo que su crecimiento es mayor, así como su popularidad. Y, por último, se encuentra la etiqueta *"Famoso"*. Son pocas las cuentas de este tipo, y aunque Twitter no tiene una cifra exacta, creemos que corresponde al 0.005% de todos los usuarios en Twitter.

Durante todo el estudio se efectuaron diferentes procesos de extracción; las primeras etapas fueron exclusivamente para encontrar patrones en los temas de interés y, una vez definidos los objetivos, se inició un proceso consistente y metódico de extracción. Al final, se obtuvieron 5.7 millones de tweets, hubo en total 13.8 millones de RT y una red de 3.2 millones de usuarios recolectados.

### *3.5. Clasificación*

Los datos que se generan de manera constante a partir de fuentes de datos se conocen como “flujos de datos”, o “data streaming” [56], y debido al hecho de que los tweets se generan en Twitter permanentemente, el tipo de datos obtenidos se engloban en esta categoría. En otras palabras, son una secuencia de elementos procesados en orden de llegada, recibidos a gran velocidad y conformados por diferentes tipos de datos. Estos datos deben procesarse de forma secuencial mediante registros o en ventanas de tiempo tal como lo hace Twitter. Las RS son un caso particular de datos de tipo *streaming*, de manera que el conjunto de entrenamiento, o datos iniciales, no necesariamente está disponibles *a-priori*, sino que se está generando constantemente, y mientras esto sucede, se debe realizar el análisis.

En términos generales, hay dos grandes fases en el procesamiento de este tipo de datos: una de almacenamiento y otra de procesamiento. La fase de almacenamiento debe tener una alta coherencia con el contenedor y no cambiar su estructura base, ya que, de lo contrario, eso ocasionaría una pérdida de información y podría tener como efecto no deseado que su tiempo de transformación se prolongara.

En cuanto a su fase de procesamiento, es necesario utilizar técnicas adecuadas a tales tipos de datos y que aborden directamente las características relacionadas con el tamaño, la velocidad de llegada, la diversidad en los datos obtenidos y las limitaciones en cuánto a recursos de almacenamiento.

En particular, se han propuesto nuevas técnicas para la fase de aprendizaje y están logrando tal auge como las presentadas en [57, 58], quienes han planteado una nueva generación de algoritmos enfocados a la denominada *stream mining* o minería de flujos de datos, que haga frente a los problemas de esta naturaleza en tiempos razonablemente cortos. Entre los algoritmos de este tipo destaca el árbol de Hoeffding (HTree) propuesto por Domingos y Hulten [59].

Aprender de los datos de flujo constante con algoritmos como HTree tiene sus ventajas, pues al estar generándose de forma dinámica y continuamente, se logra un mayor nivel de adaptabilidad y oportunidad en la obtención de los modelos correspondientes a la descripción de una realidad temporal. Entre sus desventajas se encuentran el hecho de requerir altos niveles de poder de cómputo debido a que su procesamiento que puede llegar a ser demandante.

Para que el algoritmo se llevara a la práctica se creó VFDT (Very Fast Decisión Tree) [60, 61] el cual utiliza la inecuación de Hoeffding [42] como cota de error. Domingos menciona que "la inecuación o desigualdad de Hoeffding proporciona una cota superior a la probabilidad de que la suma de las variables aleatorias se desvíe una cierta cantidad de su valor esperado". En otras palabras, la cota ayuda a determinar el número de ejemplos necesarios que garanticen que la expansión del árbol se realice de manera incremental sin afectar la precisión del mismo. Básicamente, indican que la probabilidad de que una variable aleatoria *X* se desvíe del valor esperado *E[X]* sea pequeña. Adicionalmente, la cota puede ser calculada eficientemente lo que permite su aplicación en datos cambiantes a gran velocidad, como es el caso de los tweets.

Sea $X_1, \ldots, X_n$ una variable aleatoria que puede expresarse como la suma de *N* variables aleatorias independientes idénticamente distribuidas, tal que $0 \leq Xi \leq 1$, entonces

$$X = \sum_{i=1}^{N} Xi \tag{1}$$

Dónde si $X_1, \ldots, X_n$ son variables aleatorias independientes y $a_i \leq X_i \leq b_i$ $(i = 1,2\ldots, n)$, entonces para $\theta > 0$ y $\overline{X} = \frac{1}{n}\sum_{i=1}^{N} X_i$:

$$P(\overline{X} - E[\overline{X}] > \theta) \leq e^{-} \frac{2 \cdot \theta^2}{\sum_{i=1}^{N} (a_i - b_i)^2} \tag{2}$$

Si $X_1, \ldots, X_n$ son variables aleatorias independientes que corresponden al intervalo $[a, b]$, y si $\overline{X} = \frac{1}{n} \sum_{i=1}^{N} X_i$, entonces $\theta > 0$:

$$P(E[\overline{X}] - \overline{X} > \theta) \leq e^{-} \frac{2 \cdot \theta^2}{\sum_{i=1}^{N} (a_i - b_i)^2} \tag{3}$$

Una de las grandes ventajas de VFDT es que el tiempo y la memoria necesarios para la construcción del árbol es lineal y no depende del total de casos (*tweets*) que se obtengan, sino del tamaño de *n (observaciones independientes)*.

De manera general, este algoritmo se encarga de construir un árbol de decisión de manera recursiva, reemplazando las hojas con los nodos representativos que engloban las estadísticas de los valores de cada atributo. Los nodos contienen los atributos de división y las hojas únicamente las etiquetas de la clase. Cuando se obtiene una nueva muestra se obtiene se valida el árbol completo y se evalúa cada atributo relevante en cada nodo. El algoritmo analiza cada condición posible en función de los valores de los nuevos atributos, lo que ayuda a una reconstrucción más rápida, y, en caso necesario, únicamente reemplaza los nodos necesarios.

A través de éste, y conforme se va obteniendo el flujo de datos, se construye dinámicamente el árbol de decisión, el algoritmo se presenta a continuación:

```
Permitir que HTree sea un árbol con una sola raíz (root).
```

```
Para todos los ejemplos de entrenamiento hacer
      Ordena el ejemplo en la hoja l usando HTree
      Actualiza suficientes estadísticas en l
      Incrementa nl, el número de ejemplos vistos en l
      Si nl mod min = 0 y los ejemplos vistos en l no tienen la misma clase
entonces
            Calcula Gl(Xi) por cada atributo
            Dejar Xa sea un atributo con el más alto Gl
            Dejar Xb sea un atributo con el segundo más alto Gl
            Calcula la cota de Hoeffding
            Si Xa != X0 y (Gl(Xa)-Gl(Xb) > e o e < t entonces
                  Reemplaza l con un nodo interno y divide en Xa
                  Para todas las ramas de la división hacer
                        Agrega una nueva hoja con suficientes
            estadísticas inicializadas
                  Fin Para
            Fin Si
      Fin Si
Fin Para
```

*Algoritmo 1.0 Hoeffding Tree*

Aunque el principal inconveniente es que existen variaciones que en ocasiones comprometen el coste computacional y su complejidad, esto ocurre normalmente cuando la cantidad de los atributos cambia, y no es el caso de nuestra investigación.

Además de considerar HTree usamos tres variantes del mismo (Adaptative, Option y Random); a partir de la aplicación de estas técnicas creamos de manera automatizada reglas de asociación conforme se genera este árbol incremental. Se explican a continuación las principales características de las variantes de HTree que seleccionamos. “Adaptative Tree”, como su nombre lo indica se adapta automáticamente al cambio, es tan preciso como VFDT, requiere menos memoria y no necesita una ventana de ejemplo. La variante de “Option”, son HTree regulares

que contienen nodos de opciones adicionales, su estructura única representa eficientemente árboles múltiples y mejora en que cada hoja almacena una estimación del error actual. Por último, “Random” similar a Random Forest, genera muchos árboles de clasificación sin podar y se selecciona el que tiene mejor clasificación.

Considerando lo antes planteado, adicionalmente realizamos un estudio y comparación con otros algoritmos de clasificación y aunque los datos no tienen la característica de ser incrementales, nuestro objetivo fue evaluar la precisión y el tiempo de HTree.

En esta etapa los atributos ya habían sido seleccionados y pre-procesados los datos. Inicialmente utilizamos tres conjuntos de datos que obtuvimos de la extracción en Twitter y llamamos “Tweets 1” conformado por 1,000 tweets, “Tweets 2” que contiene 10,000 tweets y por último “Tweets 3” con 100,000 tweets.

Cabe mencionar que en todas la evaluaciones del árbol de Hoeffding configuramos a VFDT con los parámetros por omisión de la herramienta *“Massive Online Analysis”* en sus siglas en inglés *MOA* [62], y de manera similar se utilizó la herramienta *Weka* para la evaluación de los algoritmos de clasificación tradicionales.

En el caso de VFDT la evaluación fue diferente a los otros, debido a que es de flujo incremental y el árbol cambia únicamente si los nuevos ejemplos son representativos.

| | ID3 | | J48 | | Naive Bayes | | Hoeffding | |
|---|---|---|---|---|---|---|---|---|
| | Precisión | Tiempo | Precisión | Tiempo | Precisión | Tiempo | Precisión | Tiempo |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tweets 1 | 90.92% | 0.89 seg | 89.1% | 1.01 seg | 90% | 0.86 seg | 91.01 | 0.75 seg |
| Tweets 2 | 87.91% | 4.19 seg | 89.21% | 3.5 seg | 84.20% | 3.87 seg | 94.40% | 2.9 seg |
| Tweets 3 | 88% | 6.19 seg | 87.19% | 5.20 seg | 87% | 5 seg | 97.10% | 5.87 seg |

*Tabla 4. Resultados de la evaluación de algoritmos de clasificación*

Como se puede observar en la tabla anterior, el árbol de Hoeffding tiene mejor precisión conforme tiene nuevos datos para evaluar, en la Figura 1 se muestran los cuatros algoritmos evaluados y en dónde cada punto en la gráfica representa los conjuntos de datos clasificados.

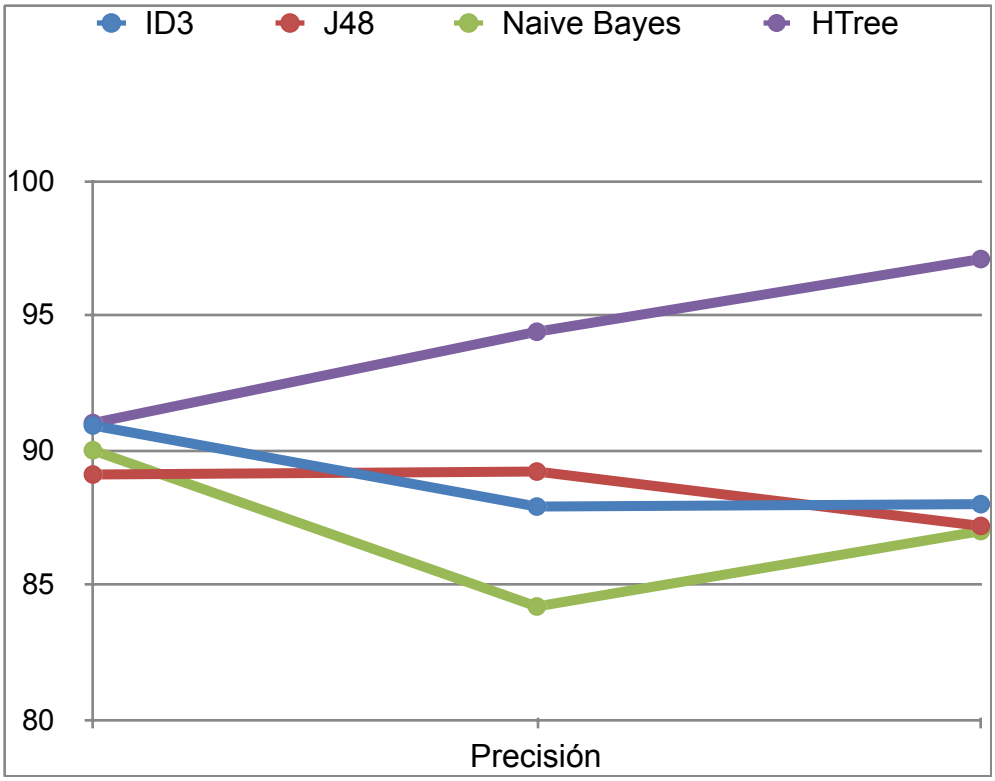


*Figura 1. Precisión de los algoritmos en los datos clasificados*

La exploración de dicho algoritmo y su selección de variantes es una de las aportaciones de la investigación realizada.

### *3.6. Reglas de asociación*

En las tareas de clasificación el modelo es visto como un árbol de caminos, dado que un nodo puede tomar diferentes rutas. Las reglas de asociación constituyen una de las técnicas más utilizadas en la minería de datos por su aplicación en diversas áreas de interés, y se definen como un conjunto de elementos y caminos que conducen a otro elemento y que pueden contener características que describen la importancia y la consecuencia en las relaciones obtenidas de un proceso de análisis mediante aprendizaje automático. Se utilizan para descubrir hechos que ocurren dentro de un determinado conjunto de datos, en nuestro caso, los tweets.

Las reglas de asociación conseguidas a través de estas técnicas expresan patrones de comportamiento en los datos en función de la aparición de valores correspondientes a dos o más atributos; debido a su forma de aplicación, se consideran algoritmos no supervisados, ya que se encargan de descubrir patrones y tendencias a partir de un conjunto de datos, en este caso, tweets relacionados con los HT.

Se trata de una descripción de los atributos y sus valores correspondientes, donde existe un antecedente y sus consecuentes. Se integraron en el método de predicción para encontrar caminos descriptivos de los HT que se convirtieron en TT, y sirvió para aprender de los temas posicionados y los factores que los volvieron exitosos. De esta manera fue posible, en una primera instancia, dado un conjunto de tweets y usando el árbol de Hoeffding, descubrir y describir a continuación, a través de las reglas de asociación, lo averiguado previamente que, a la postre, no es más que la descripción de las características que pueden predecir la ocurrencia del RT, basado en la presencia de los elementos recomendados y, por ende, en una posible viralización del HT que contiene dichos tweets.

Existen algunos métodos clásicos para la extracción de las reglas de asociación, entre los algoritmos más populares son Eclat [63, 64], FP-Growth [65] y A-priori [66, 67]. El algoritmo Eclat es similar a A-priori, para cada elemento (item) almacena en una lista la transacción a la que pertenece cada *item*, de esta manera se reduce el

tiempo de cómputo. FP-Growth genera una representación reducida de la base de datos mediante árboles, como requisito los *items* deben estar ordenados.

Por último, el algoritmo A-priori es uno de los más comunes y el que usaremos, se basa la frecuencia de sus *items* y dónde todos sus subconjuntos también deben serlo. Los autores sugieren una notación que permite definir la regla de asociación como una implicación de la forma *X* => *Y*, en donde *X* (antecedente) e *Y* (consecuente) son conjuntos de *elementos,* en este caso de tweets.

Se pueden definir dos medidas de interés para cada regla: el soporte o cobertura y la confianza. El primero expresa el porcentaje o fracción de registros de que satisfacen la unión de los elementos del antecedente y del consecuente de la regla. Por otro lado, la confianza es la medida de efectividad de la regla, representada mediante el porcentaje de casos en los que dado *X* se verifica la implicación. De acuerdo con Neves et al. [68] se recomienda que los parámetros de entrada del algoritmo se definan bajo un valor de soporte y un valor elevado para la confianza. Considerando la recomendación, en primer lugar se genera una gran cantidad de reglas y posteriormente se validan a través de la medida de confianza.

Las reglas obtenidas se generaron a partir del conjunto de datos recopilados de los tweets correspondientes a las tres categorías de interés (*Deportes, Espectáculos y Política*), y se hicieron basándonos en los nueve atributos antes mencionados en la tabla 5.

Las características del algoritmo *a-priori* son las siguientes: soporte mínimo de 0.45 y confianza mínima de 0.9, como resultado presentamos las cinco mejores reglas y que más se repetían con cada conjunto de datos a analizar. Cabe destacar que las reglas obtenidas durante esta etapa lograron que los tweets publicados tuvieran mayor alcance en función de los RT, en comparación con aquellos que fueron publicados sin seguir las recomendaciones obtenidas mediante las reglas de asociación.

| ID | Reglas de asociación |
|---|---|
| 1 | **Category** => e AND **totalWords** => 2 AND **seasonal** => 0 AND **days** => Sat OR (**days** => Fri AND **hourDay** => 4) OR **days** => Sun |
| 2 | **Category** => p AND **totalWords** => 1 AND **seasonal** => 0 AND **uInfluence** => Outstanding1 |
| 3 | **Category** => e AND **totalWords** => 2 AND (**days** => Sat AND **hourDay** => 3) OR (**days** => Sat AND **hourDay** => 4) AND (**uInfluence** => Outstanding1 OR **uInfluence** => Famous) AND **additionalContent** => 1 |
| 4 | **Category** => p AND **totalWords** => 2 AND (**days** => Sat AND **hourDay** => 4) AND (**uInfluence** => Outstanding1) AND **siteType** => portalA |
| 5 | **Category** => d AND **totalWords** => 2 AND **seasonal** => 1 AND (**days** => Sat AND **hourDay** => 3) OR (**days** => Sun AND **hourDay** => 4) AND (**uInfluence** => Outstanding2 OR **uInfluence** => Outstanding1 OR **uInfluence** => Famous) AND **additionalContent** => 1 |

*Tabla 5. Ejemplos de reglas de asociación*

La primer regla representa que para que un HT de la categoría *Espectáculos* tenga mayor oportunidad de convertirse en TT debe publicarse preferentemente los sábados o, en otro caso los viernes entre las 15:00 h a las 19:00 h, o bien los domingos. La regla dos se interpreta que para la categoría *Política* se recomienda un HT de una sola palabra, sin importar que no sea estacional y mencionando a usuarios con influencia *Outstanding1,* es decir, que su rango de seguidores se encuentre entre 10,0001 a 100,000.

En cuanto la categoría *Espectáculos* se presentan las recomendaciones en la regla *tres,* mediante la cuál se recomienda un HT de dos palabras concatenadas, publicar los días Sábado entre las 15:00 h a las 19:00 h, y mencionar a usuarios

*Outstanding1* o *Famosos,* además de incluir contenido en el tweet sin importar el tipo.

En *la fila 4 correspondiente a la tabla* se presentan las recomendaciones para la publicación de la categoría *Política,* en dónde se debe publicar un HT de dos palabras concatenadas, publicar de preferencia el día sábado entre las 15:00 h a las 19:00 h, mencionar o lograr que usuarios de tipo *Outstanding1* hagan RT y publicar contenido de tipo *portalA,* es decir, contenido serio.

Y por último, en *fila 5* describe para la categoría *Deportes* el HT debe ser de dos palabras concatenadas, en un periodo estacional, y los días sábado o domingos, incluyendo en el mensaje o logrando RT de usuarios de tipo *Outstanding1*, *Outstanding2* o *Famous,* además de la inclusión de contenido adicional.

Las reglas anteriores pueden cambiar conforme el árbol continúa aprendiendo, son únicamente las cinco que obtuvimos con mayor frecuencia durante nuestro estudio.

Si bien existen entornos de minería de datos que proveen algoritmos para generar reglas de asociación, nuestra propuesta va más allá, debido a que se realiza una extracción, pre-procesamiento y clasificación de datos incrementales, los cuáles brindan la posibilidad de mejorar y obtener nuevas reglas de asociación dependiendo de los TT y tweets extraídos. Es decir, se adapta al entorno, dando como resultado un método de predicción que se puede ajustar a los intereses de los usuarios.

# CAPÍTULO IV: Arquitectura de sistemas multi-agentes

El método de predicción que proponemos está conformado por los procedimientos mencionados en el capítulo anterior, a medida que se obtuvieron los tweets las tareas se tornaron complicadas. Por lo que fue preciso encontrar una arquitectura capaz de resolver problemas de escalabilidad, automatización y distribución, y dadas las características de un sistema multiagentes (SMA) fue la solución adecuada.

El objetivo de este capitulo es presentar la arquitectura propuesta basada en SMA y los elementos que la conforman, además de sus principales ventajas y las interacciones entre los agentes que permiten automatizar nuestro método de predicción.

Otro beneficio de la arquitectura SMA es su escalabilidad, es decir, podemos agregar nuevos agentes y capacidades sin afectar a los agentes ya existentes, y favorecer la solución de las diferentes tareas.

Un SMA consiste, así, en un conjunto de agentes que interactúan entre sí, y que además están constituidos por sus creencias, deseos e intenciones, para los cuales fueron diseñados. Los agentes tienen la capacidad de responder de manera oportuna a los cambios en el ambiente, efectuar un control de sus acciones, comunicarse con otros agentes y adaptar su comportamiento en función de la "experiencia" adquirida. Shoham [69] define a un agente como "*Una entidad de software la cual funciona continuamente y autónomamente en un ambiente en particular, a menudo habitada por otros agentes y procesos*".

Adicionalmente Maes [70] los define como "*sistemas computacionales que habitan en entornos dinámicos complejos, perciben y actúan de forma autónoma en ese entorno, realizando un conjunto de tareas, y cumpliendo los objetivos para los cuales han sido diseñados*".

De acuerdo con los autores [71, 72] un SMA permite encontrar respuestas a problemas que están más allá de las capacidades individuales o del conocimiento de cada entidad. La cooperación dentro de un SMA puede realizarse de tres formas:

1. Por diseño explícito, es decir, se asignan los comportamientos de cada agente.
2. Por evolución, que consiste en la cooperación de los agentes que cambian a través de un comportamiento evolutivo.
3. Por adaptación, donde los agentes, de manera individual, aprender a cooperar.

Considerando lo antes mencionado, optar por el uso de SMA como arquitectura resultó ser la elección más conveniente debido a que los agentes a menudo se ocupan de resolver aplicaciones complejas de sistemas distribuidos y minería de datos, y sus funcionalidades pueden extenderse a nuevos agentes o inclusive modificar sus objetivos.

La interacción lograda entre los agentes es casi de manera natural mediante la cooperación, la coordinación, la negociación y la comunicación en un ambiente determinado que conceden una arquitectura robusta, escalable, sólida y con la autonomía necesaria para la solución de problemas complejos, como es nuestro caso de estudio.

A continuación se presenta la arquitectura propuesta y los agentes que la componen.

### *4.1. Arquitectura.*

En este apartado presentamos una arquitectura distribuida que se adapta al ambiente sin la necesidad de la intervención humana y tiene la capacidad de tomar decisiones considerando los posibles cambios en su entorno. Para ello, optamos por una arquitectura deliberativa por su funcionamiento basado en el razonamiento y que además posee una representación interna del mundo. Rao y Georgeff [73, 74] propusieron una estructura lógica que fundamenta su modelo en BDI Beliefs -

Desires - Intentions (*Creencias - Deseos - Intenciones*). Su éxito se basa en simplificar lo complejo del comportamiento humano y su lógica, en una instancia conformada por agentes, metas, planes, creencias, deseos e intenciones.

Se describen a continuación sus tres elementos principales:

a) Creencias (***B**eliefs*). Representan el conocimiento que tiene cada agente sobre el entorno y determinan su visión del mundo; es un componente informativo.
b) Deseo (***D**esires*). Es lo que el agente desea conseguir y que por lo tanto, es una de sus prioridades; están representados por sus motivaciones.
c) Intenciones (***I**ntentions*). Modelan los objetivos que hay que alcanzar y existe un compromiso por su parte para conseguirlos.

Considerando la notación del concepto de agente BDI podemos explicarlo como:

$\theta$ es el conjunto que describe el entorno del agente
Siendo $T(\theta)$ el conjunto de atributos $\{\tau_1\tau_2, \ldots, \tau_n\}$ en el que son expresados
$\tau_n$ representa los atributos definidos en el capítulo anterior, y los cuáles son necesarios para la generación de recomendaciones.

Teniendo en cuenta que $T(\theta)$ incluye los atributos para determinar qué reglas se generan a partir de una categoría es posible expresarlo como:

$$T(\theta) = \{\tau_1 =' resultType', \tau_2 =' hourDay', \tau_3 =' additionalContent', \ldots, \tau_n\}$$

En dónde, $\tau_1$ corresponde al atributo que describe el tipo de tweet, $\tau_2$ la hora del día en que fue publicado y $\tau_3$ si el tweet tiene contenido adicional. Por lo que cada $\tau_n$ representa un atributo y $T(\theta)$ el conjunto de ellos.

Por ejemplo, si definimos a la categoría *deportes* como una creencia tenemos,

$$deportes = (\tau_1 =' resultType', \tau_2 =' hourDay', \tau_3 =' additionalContent', \ldots, \tau_n)$$

Cada una de las creencias expresadas se almacenan de manera aislada por cada agente, además de sus intenciones y deseos, en la figura 2 podemos ver la representación más simple.

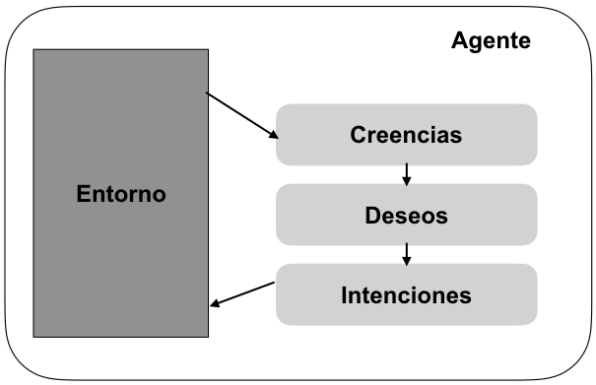


*Figura 2. Arquitectura BDI*

Dividimos las tareas del método en bloques que a su vez son resueltas por uno de los tres agentes que proponemos: *Coordinator, Crawler* y *Predictor,*

La arquitectura propuesta divide las tareas del método en bloques que a su vez son resueltas por uno de los tres agentes que proponemos: *Coordinator, Crawler* y *Predictor.* Se definieron las interacciones entre ellos de manera que el agente *Coordinator* sea el único que conozca en qué etapa se encuentran los otros dos agentes, es el encargado de organizar las tareas correspondientes a la arquitectura propuesta.

Cada agente tiene asignado un conjunto de tareas que únicamente él puede realizar, de manera que fue necesario contar con un mecanismo que proporcione los elementos básicos para una comunicación adecuada entre los diferentes agentes y sus actividades.

Para ello nos respaldamos en el enfoque de agentes y artefactos (A&A) [75]. Los artefactos representan los recursos que los agentes pueden utilizar y compartir de manera dinámica para apoyarse en sus actividades individuales y colectivas. Adicionalmente el uso de CArtAgO [76] nos permite crear espacios de trabajo que se conciben como un conjunto dinámico de elementos que proporcionan una comunicación con los agentes, los cuales pueden participar en uno o más espacios de trabajo, en el caso abordado en el presente trabajo únicamente se utiliza un espacio.

Un elemento adicional importante para que la arquitectura funcione es la manera en que los agentes se comunican y así establecer las solicitudes y respuestas que cada agente es capaz de realizar. Esta interpretación del lenguaje de los agentes se efectúa mediante la notación lógica propuesta originalmente por Rao [77, 78, 79] definida como *AgentSpeak(L);* es un lenguaje de programación basado en lógica restringida de primer orden que consiste de acciones y eventos.

Los elementos descritos componen la arquitectura propuesta para que el método de predicción se lleve a cabo de manera dinámica y ordenada, en donde cada componente podrá interactuar de una forma transparente, se presenta de manera general en la Figura 3.

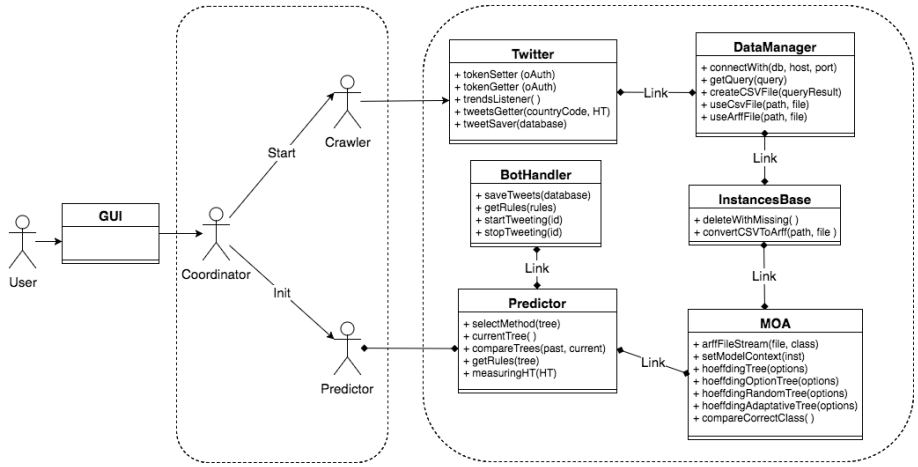


*Figura 3. Arquitectura del método de predicción*

Para que la plataforma BDI funcione adecuadamente es necesario definir el ambiente, las limitaciones y el alcance de cada agente, a continuación los describimos a detalle.

### *4.2. Definición de los agentes*

Para la presente investigación el objetivo fue adquirir conocimiento de acuerdo a las categorías definidas inicialmente (deportes, espectáculos y política) y a partir de ello obtener reglas de asociación que a su vez se convierten en recomendaciones que pueden ser integradas al contenido del tweet. Su capacidad de acción depende de sus tareas asignadas y la manera en que se comuniquen con otros agentes para lograr la meta en común.

Las tareas principales que los agentes deben realizar son: extracción de tweets, limpieza en los datos, aprendizaje mediante el árbol de Hoeffding y la generación de reglas de recomendación. Se presenta a continuación las capacidades y funciones que realiza cada agente.

**a) Agente *Coordinator* (Co).**

- *Función:* Se encarga de controlar las acciones del método de predicción y tiene la tarea de enviar los actos de habla a los agentes *Crawler* y *Predictor.* Cuando tales agentes terminan sus tareas el agente *Coordinator* conoce las reglas de inducción entregadas por *Predictor,* y es en ese momento cuando estas reglas se presentan como recomendaciones para el usuario por medio de una aplicación Web.
- *Capacidades:* Es el único que conoce en qué fase se encuentra el flujo de trabajo, además de recibir las solicitudes del usuario desde una plataforma Web, de manera que el agente decide en qué momento se inician y finalizan las tareas relacionadas con el método de predicción.

**b) Agente Crawler (Cr)**

- *Función:* Una vez que el agente *Coordinator* realiza una petición al agente *Crawler* éste inicia la extracción de nuevos tweets y además se conecta a una base de datos donde, mediante una consulta, obtiene los tweets relacionados con la categoría solicitada por el *Coordinator.*
- *Capacidades:* Además de la extracción puede generar archivos de tipo atributo-relación (.arff), que consisten en la declaración de las variables y la descripción de su tipo.

**c) Agente Predictor (Pr)**

- *Función:* Su tarea principal es el análisis de los tweets que previamente fueron extraídos y almacenados por el agente *Crawler*. El análisis consiste en realizar una clasificación mediante el árbol de Hoeffding y utilizar tres variantes del algoritmo original, como son Random Tree, Adaptative Tree y Option Tree. Una vez realizada la clasificación genera las reglas definidas por el árbol seleccionado y se las envía al agente *Coordinator*.
- *Capacidades:* El agente *Predictor* compara los modelos generados y selecciona el que logra una mejor clasificación en un periodo de tiempo menor.

### *4.3. Artefactos*

Los artefactos representan componentes pasivos empleados por los agentes para conseguir sus metas; éstos proporcionan una función dividida en diferentes operaciones que los agentes pueden emplear al interactuar con el artefacto. Para ello es necesario crear espacios de trabajo, que son contenedores conceptuales de agentes y artefactos que permiten definir la topología del entorno.

Los artefactos incluidos en la arquitectura son:

a) *Twitter*. El artefacto provee las operaciones para que el agente *Crawler* extraiga los tweets mediante el uso de la API de Twitter. Sus operaciones son:

- *tokenSetter*. asigna y guardar en la base de datos las claves de autenticación proporcionadas por Twitter;
- tokenGetter: obtiene las claves para la autenticación con Twitter;
- trendsListener; adquiere las tendencias del momento;
- tweetsGetter: extrae las publicaciones que tienen en su contenido el HT solicitado por el agente *Coordinator;*
- tweetsSaver: almacena los tweets en una base de datos no relacional para mantener la estructura base del contenido.

b) *DataManager*. Encapsula algunos métodos de Weka utilizados para la selección de atributos y la generación de archivos .arff. Además, permite al agente *Crawler* conectarse a la BD para hacer consultas.

- connectWith y getQuery: estas operaciones se encargan de trabajar con la base de datos seleccionada por el usuario a través del agente *Coordinator* y hacer consultas, respectivamente;
- createCSVFile: si los tweets se obtienen desde una consulta se crea un archivo en formato .csv (Comma-Separated Values), con los atributos correspondientes;
- useCSVFile y useArffFile. estas operaciones se encargan de seleccionar el archivo que contiene los tweets para su evaluación y clasificación.

c) *InstancesBase.* Este artefacto encapsula los métodos *deleteWithMissing* y *convertCSVToArff* pertenecientes a Weka.

d) *MOA*. Encapsula los métodos de MOA, entre ellos el árbol de Hoeffding, *compareCorrectClass* y *arffFileStream*;

e) *El agente Predictor.* Utiliza algunos métodos de MOA para realizar las siguientes operaciones:

- compareTrees: compara los resultados de clasificación y el tiempo de ejecución de las variantes del árbol de Hoeffding;
- selectMethod: selecciona el método del árbol de Hoeffding con mejor clasificación y en menor tiempo;
- currentTree: obtiene los nodos del árbol seleccionado;
- getRules: la operación se encarga de obtener las reglas de asociación del árbol seleccionado;

- measureHT: una vez generadas y aplicadas las reglas de asociación por el usuario, esta operación analiza cada quince minutos el comportamiento del tweet que hay que posicionar. Mediante esta operación se obtiene el total de RTs, favoritos y usuarios que se unen a la conversación de la publicación.

f) *BotHandler.* La tarea principal del artefacto es automatizar el proceso de publicación de tweets creados por el usuario.

- saveTweets: guarda los tweets creados por el usuario en una base de datos definida;
- getRules: por medio del agente *Coordinator* y el panel de control, el usuario visualiza las diferentes recomendaciones para crear el contenido de sus tweets;
- startTweeting y stopTweeting: las operaciones se encargan de iniciar y detener el proceso de publicación automatizada.

Cada uno de los artefactos está compuesto con las operaciones necesarias para que los agentes *BDI* logren las metas que se les han sido asignadas. Estos son los elementos que constituyen la base de la arquitectura necesaria para que el método de predicción propuesto y para que el diseño de la misma sea viable utilizamos un framework de creación de artefactos denominado CArtAgO, el cuál provee de una infraestructura que hace posible trabajar con A&A para modelar y plantear entornos de trabajo colaborativos donde lo agentes son entidades y los artefactos son sus recursos, se presenta a continuación la arquitectura.

Para que lo antes descrito sea más claro nos apoyamos de un lenguaje de visualización, especificación y construcción como es UML (Unified Modeling Language) [80], siendo más específicos mediante un diagrama de secuencia (Ver Figura 4), que muestre el comportamiento de los agentes, artefactos y sus respectivos mensajes.

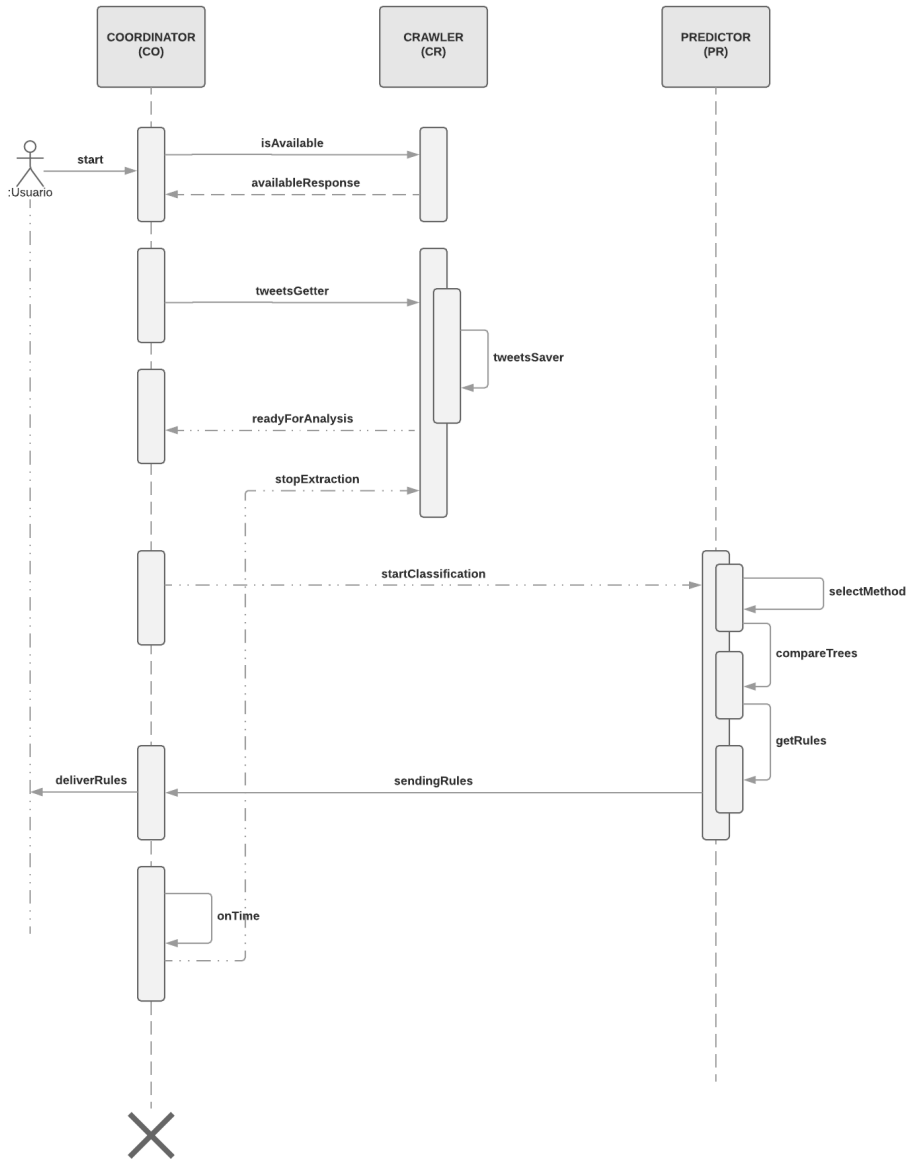


*Figura 4. Diagrama de secuencia*

Como puede observarse cuando el usuario fuera del ambiente de trabajo de los agentes realiza una solicitud para obtener recomendaciones de posicionamiento del HT inicia el flujo del método de predicción, y es únicamente el agente *Coordinator* quien recibe peticiones por parte del usuario y a su vez encarga de transmitir las instrucciones a los otros dos agentes restantes. Se describe a detalle:

1. El usuario mediante la aplicación Web solicita la extracción (*start*) de un tema en particular.
2. El agente *Coordinator* obtiene la información proporcionada por el usuario correspondiente a las especificaciones de la extracción.
3. *Coordinator* valida que el agente *Crawler* esté disponible para realizar la tarea.
4. El agente *Coordinator* envía la solicitud de extracción a *Crawler* con sus parámetros correspondientes como son: la categoría del tema, inicio y fin de la extracción.
5. El agente *Crawler* inicia la extracción de nuevos tweets.
6. El agente *Crawler* se conecta a una base de datos donde obtiene los tweets históricos relacionados con la categoría solicitada por el agente *Coordinator.*
7. Cada vez que el agente *Crawler* ha extraído mil tweets le envía un mensaje al agente *Coordinator* para indicarle que está listo un bloque para ser analizados.
8. *Coordinator* le solicita al agente *Predictor* que inicie la clasificación con el bloque indicado por el agente Crawler.
9. El agente *Predictor* aplica cuatro modelos de aprendizaje mediante el árbol de Hoeffding y sus variantes Random Tree, Adaptative Tree y Option Tree.
10. En cada iteración el agente *Predictor* crea un nuevo modelo mediante al algoritmo, el cuál se encarga de valorar si es necesario reestructurar.
11. El agente Predictor selecciona el modelo que logra una mejor clasificación y en un tiempo menor.
12. Una vez realizada la clasificación el agente *Predictor* genera las reglas definidas por el modelo seleccionado y se las envía al agente *Coordinator*.
13. El agente *Coordinator* presenta al usuario mediante la plataforma web las reglas obtenidas de la clasificación.
14. *Coordinator* verifica si la fecha y hora del proceso de extracción es vigente.
15. En caso de llegar al tiempo de expiración le envía un mensaje al agente *Crawler* para que detenga la extracción.
16. El agente *Coordinator* continúa pendiente de nuevas posibles solicitudes para iniciar el proceso antes mencionado.

La finalidad de permitir que por cada solicitud se definan los parámetros es que el método se pueda ajustar de cierta forma al requerimiento del usuario, por lo tanto cada solicitud es diferente pero se mantiene el flujo de trabajo.

En el siguiente capitulo se explican los experimentos correspondientes a cada categoría y las reglas de decisión generadas a partir del método de predicción.

# CAPÍTULO V: Experimentos y Resultados

En el presente capítulo explicamos los experimentos realizados en las diferentes etapas del descubrimiento de los atributos hasta llegar a un porcentaje de clasificación adecuado y a su vez obteniendo mejores resultados con el algoritmo de HTree y sus correspondientes variantes.

Los experimentos efectuados se caracterizan por lograr los siguientes objetivos:

* Demostrar la viabilidad del algoritmo de HTree
* Indicar cuáles de las variantes de HTree es mejor y en qué condiciones
* Demostrar que la arquitectura de sistemas multiagentes soluciona el problema de autonomía y coordinación
* Probar la validez del método de predicción propuesto
* Mostrar evidencias del uso de las reglas de decisión en algunos tweets publicados

Adicionalmente a la investigación logramos crear una aplicación Web que denominamos *TweetPop* en su fase prototipo que facilitó la evaluación de las reglas y además que un grupo de usuarios pusiera en práctica la arquitectura SMA sin tener conocimiento de la misma.

Para una mejor comprensión de los experimentos hemos dividido en secciones en las que detallamos cómo se conforman los conjuntos de datos disponibles para la evaluación de HTree, posteriormente la comparación con las variantes de HTree presentando el tiempo de ejecución y la precisión de clasificación. Posteriormente las reglas de asociación generadas a partir de los árboles generados y por último el desarrollo tecnológico originado de la presente investigación.

### *5.1. Conjuntos de datos*

Aplicamos el algoritmo de HTree y sus variantes (HTree Option, HTree Adaptative y HTree Random) a cuatro conjuntos de datos conformados por distintos números de instancias (250, 500, 750 y 10,000) correspondientes a cada categoría. Estas instancias se formaron mediante la extracción de los tweets que tuvieron en su contenido uno de los HT que previamente definimos de acuerdo a nuestras categorías de interés.

Nuestro primer conjunto de datos (C1) únicamente tiene los atributos que brinda Twitter, como son: tipo del tweet, hora de publicación, día de publicación, contenido adicional, total de seguidores y total de RT. Sin embargo, debido a los resultados alcanzados no eran atributos suficientes para una clasificación adecuada, por lo que decidimos que para el segundo conjunto de datos (C2) se incluyeran el atributo categoría y que el atributo de hora de publicación sea un valor calculado e indique el intervalo de hora en que fue publicado el tweet. Hasta el momento no eran suficientes los atributos para que HTree o sus variantes lograran una progreso considerable en la precisión de la clasificación.

Fue entonces que nuestra investigación nos llevó a determinar los atributos que llamamos como derivados, debido a que Twitter no los proporciona directamente, son atributos que calculamos y después son almacenados.

Nuestra investigación sugiere que con los atributos derivados la precisión mejora notablemente; es por ello que creamos cuatro conjuntos de datos más para confirmar nuestra propuesta.

En el tercer conjunto de datos (C3) añadimos a los atributos antes mencionados nuestro primer atributo calculado llamado “tamaño de la palabra” (**`wordLenght`**) que es un valor numérico correspondiente al total de las palabras concatenadas que forman el HT.

Otro atributo propuesto se deriva de lo que observamos durante nuestra investigación, y se debe a que existen HT que se repiten año con año y que además en su mayoría se convierten en TT, como son: #SuperTazón, #LigaMX y #Grammys por mencionar algunos, dicho atributo lo llamamos “estacional” (**`seasonal`**), además integramos en el cuarto conjunto de datos (C4) un atributo adicional llamado “tipo del sitio” (**`siteType`**) el cuál almacena la clasificación en caso de que el tweet tenga incluido algún enlace.

Por último, en el quinto conjunto (C5) incorporamos el atributo que asigna la “clasificación del usuario” (**`uInfluence`**) definido por el número de usuarios que lo siguen.

Debido a tuvimos cinco conjuntos de datos por cada categoría y a su vez divididos por el número de instancias (250, 500, 750 y 10,000) fue indispensable definir una nomenclatura que nos ayudara a reconocer cuáles datos se estaban evaluando y apoyara a la comprensión de los resultados.

Para su identificación cada subconjunto de datos tiene una etiqueta que se compone de la primer letra en mayúsculas perteneciente a la categoría que se evalúa, posteriormente el identificador del conjunto de datos que fueron descritos anteriormente y para finalizar separado con un guión medio el número de instancias que lo componen.

Por ejemplo, si evaluamos las 750 instancias para la categoría deportes del conjunto de datos que incluye los atributos iniciales y además el “tamaño de la palabra” (C3) la etiqueta del archivo a evaluar es DC3-750, y en caso de evaluar el mismo número de instancia y categoría pero del conjunto de datos que tiene incluidos los atributos ofrecidos por Twitter y además los atributos derivados la etiqueta que le corresponde es DC5-750.

En el siguiente apartado se explica la configuración de los experimentos y así prevenir que los resultados se vea afectados por aspectos técnicos y no por la aplicación del algoritmo.

### *5.2. Configuración*

Comparamos los resultados conseguidos utilizando el algoritmo de HTree y sus variantes, en términos de precisión, y tiempo de ejecución, y para asegurarnos que la clasificación no se afectara por la configuración evaluamos a todos los algoritmos de la misma manera.

Para calcular el tiempo de ejecución nos aseguramos que todos los experimentos se realizaran en un mismo ambiente de trabajo y exclusivamente ejecutándose la clasificación para que ninguna tarea externa afectara en la terminación de la misma. El ambiente de trabajo utilizado fue una computadora Mac Book Pro con un procesador Intel Core i5 / 2.9 GHz y memoria de 16 GB 1867 MHz DDR3.

Relacionado a los parámetros de evaluación se utilizó el algoritmo ganancia de información para el proceso de división y una validación cruzada de diez veces. En la fase de validación usamos *prequencial* [81, 82] en la cuál cada instancia es evaluada antes de que el algoritmo entrene con ella, es recomendaba cuándo se trabaja con flujos de datos continuos como es nuestro caso de estudio. De acuerdo con los autores [83] cada instancia se puede utilizar para comprobar el modelo antes de usarlo para el entrenamiento; *prequencial* es un esquema utilizado para intercalar las fases de prueba y entrenamiento en el proceso de clasificación, es decir, nos aseguramos que todas las instancias son utilizadas una vez para entrenar y otra para su clasificación.

## *5.3. Resultados del experimento*

Los resultados del tiempo de ejecución alcanzados por los experimentos se resumen en las siguientes figuras, únicamente presentamos aquellos que corresponden a 10,000 instancias y los resultados restantes se muestran en el Anexo III.

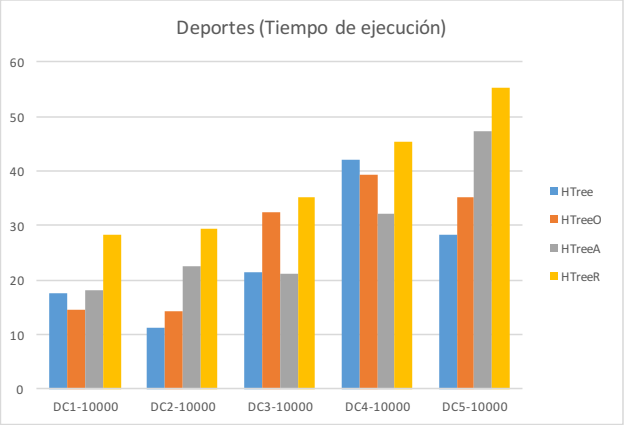


*Figura 5. Resultados del tiempo de ejecución para deportes*

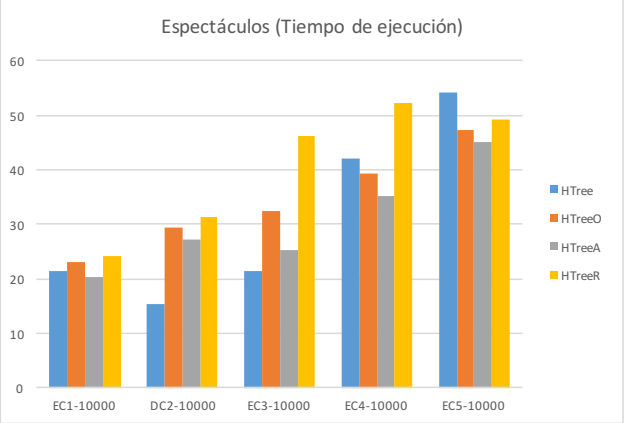


*Figura 6. Resultados del tiempo de ejecución para espectáculos*

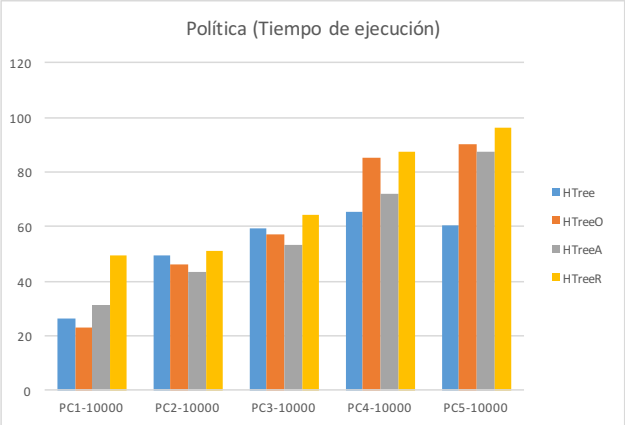


*Figura 7. Resultados del tiempo de ejecución para política*

Como puede observarse el algoritmo HTreeR es el peor caso en todas las ejecuciones con excepción del conjunto EC5-10000 en dónde obtuvo un tiempo de 49.21 segundos y el algoritmo HTree 54.1 segundos, sin embargo en las demás corridas su tiempo inclusive llegó a 96.3 segundos.

Por otro lado, los algoritmos con un tiempo de ejecución menor fueron HTree y HTreeA. Los resultados confirman el fundamento del algoritmo de HTreeA, es decir, que se ajusta a los flujos de datos que cambian con el tiempo de manera que sus parámetros no se modifican en cada ejecución, lo que ayuda a la reducción de la duración del proceso de clasificación.

Además todas las estadísticas relevantes de las instancias de HTreeA se mantienen en los nodos, por lo tanto no hay necesidad de una ventana adicional para almacenar los nuevos ejemplos dando como resultado una reducción sustancial en el consumo de memoria y el tiempo de ejecución.

Respecto a la experimentación para conocer la precisión de la clasificación correspondiente a cada algoritmo seguimos el mismo criterio de evaluación. Primero consideramos el conjunto de datos C1 hasta llegar a C5 en cada categoría, y además el número de instancias antes mencionadas.

Juzgando por los resultados de las figuras 8 a 10 el algoritmo HTreeR tuvo la peor clasificación en todas las categorías, inclusive por debajo del 70% de precisión, y si a esto agregamos el tiempo de ejecución presentado con anterioridad podemos concluir que es un algoritmo insuficiente al menos para nuestro caso de estudio. Sin importar la categoría o número de instancias la exactitud no aumentó lo suficiente, inclusive para el conjunto de datos vinculados a espectáculos (Figura 9) tuvo un descenso en la precisión.

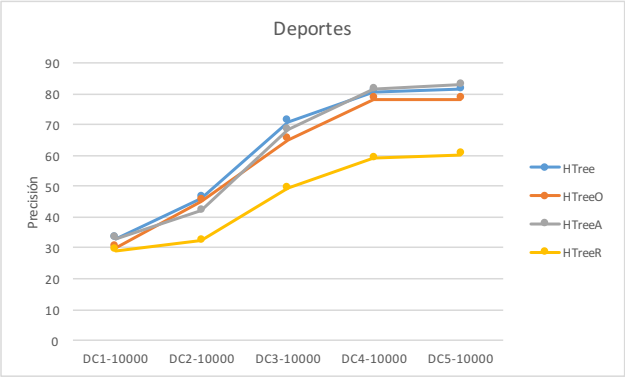


*Figura 8. Resultados de la precisión para deportes*

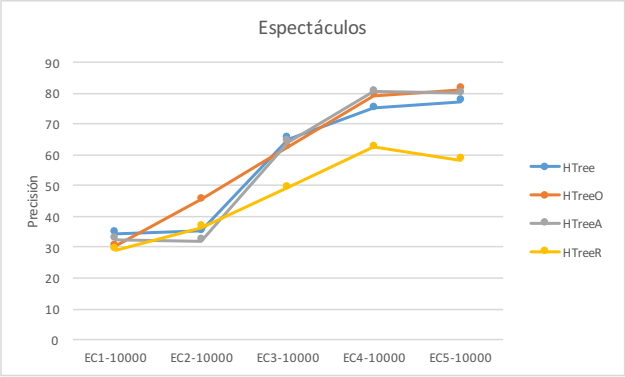


*Figura 9. Resultados de la precisión para espectáculos*

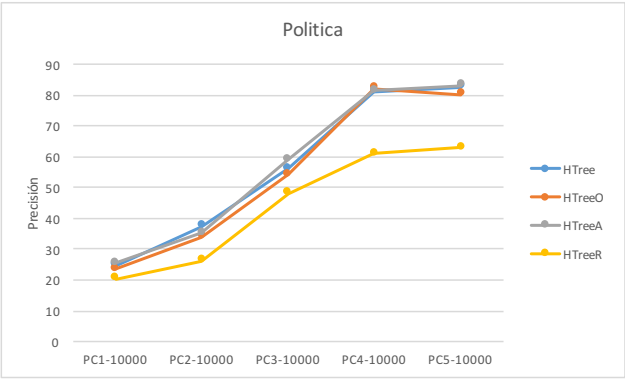


*Figura 10. Resultados de la precisión para política*

Con los resultados presentados confirmamos que en las primeras etapas de evaluación los atributos recabados de Twitter no eran suficientes y que la inclusión de en todas las categorías de los atributos propuestos a partir del conjunto de datos C4 la exactitud tuvo un crecimiento notable, con excepción del algoritmo HTreeR. Sin embargo este comportamiento es comprensible por la manera en que HTreeR divide los nodos.

En el caso de HTree y HTreeA la precisión es similar entre ellos al menos para este caso de estudio, y aunado al tiempo de ejecución los consideramos los mejores algoritmos de clasificación y generación de reglas de asociación. Además, la variante HTreeA tiene como ventaja que los árboles alternos son creados tan pronto se detecta un cambio, si HTreeA detecta que un árbol previo tiene menor precisión que uno actual lo reemplaza sin tener que esperar un número de instancias, lo que agiliza la generación de una nueva estructura.

Anteriormente los algoritmos HTree y HTreeA han sido evaluados y comparados por otros autores [84, 85, 86] dando resultados similares en cuanto a tiempo de ejecución y precisión, a diferencia que los conjuntos de datos que los autores presentaron en gran parte fueron datos artificiales y no se obtuvieron como en nuestro caso de una red social la cuál tiene una complejidad adicional, y esa es la variabilidad.

Una vez que hemos seleccionado a HTree y HTreeA como los algoritmos con mejores resultados para nuestro caso de estudio evaluamos a partir de los árboles generados las reglas de asociación de cada categoría. Una vez generadas las reglas pusimos a prueba el método de predicción para así evaluar su eficacia en cada categoría.

### *5.4. Reglas de asociación*

Examinamos la efectividad de las reglas o estrategias en un entorno cambiante como Twitter, para ello las separamos por categorías y de esta forma identificar que inclusión de atributos logró mejores resultados dependiendo si fue deportes, espectáculos y política.

La valoración de las reglas se basa en el número de RT alcanzados y el número de usuarios que interactuaron con la publicación, a mayor número de RT determinamos que su importancia es mayor. Algunas de las estrategias de publicación se comportaron mejor que otras e inclusive se reincidieron entre las recomendaciones del algoritmo, principalmente cuando la estructura del árbol ya no tenía cambios relevantes, por lo que no se obtuvieron cambios en los atributos recomendados. A continuación presentamos las reglas más importantes.


*Figura 11. Resultado de una recomendación.*

En la Figura 11 pusimos a prueba la recomendación de publicar el día sábado a las 17:00hrs y con el HT conformado por dos palabras concatenadas; lo que deseamos destacar es que un usuario con etiqueta lingüística “Destacado I” al compartirlo son sus seguidores reforzó el interés en el mensaje, logrando un incremento de 15 RT provenientes de su lista de seguidores.

Dicha regla de asociación es presentada por el agente *Predictor* de la siguiente manera:

```
day=Sat category=D dayInterval=4 wordLenght=2 ==> trending=Yes
```

Una de las recomendaciones de la categoría deportes que tuvo una consecuencia positiva por medio de la acción de hacer favorito el mensaje y no por medio del RT fue el siguiente (Figura 12).


*Figura 12. Resultado de recomendación aplicada a la categoría Deportes.*

Aunque la acción de hacer favorito un mensaje no tiene como objetivo alcanzar nuevos seguidores como es el caso de efectuar un RT, en nuestro caso logró ampliar el número de usuarios y por ende aquellos con etiqueta lingüística “Destacado I”,

“Destacado II” e inclusive “Famoso”. Fue tal el interés en el mensaje que incluso Twitter clasificó a la publicación como un “Momento” dentro de la red, esto significa que logró un impacto importante y que se debe destacar para continuar su divulgación. Su manera de presentarse es:

```
additionalContent=1 category=D wordLenght=2 siteType=image ==>
trending=Yes
```

La regla más simple que podemos obtener y que tiene menos posibilidades de conseguir RT es la que no tiene elementos adicionales, sin embargo el atributo derivado llamado **`seasonal`** adiciona la oportunidad de que la publicación sea descubierta como es el caso presentado en la siguiente figura.


*Figura 13. Resultado de recomendación con el atributo seasonal.*

La regla anteriormente presentada se explica como
category=D wordLenght=3 ==> trending=Yes

Sin embargo, otro caso de una publicación con la recomendación (Figura 14) sin inclusión de elementos y que en este caso tampoco era ocasional se efectuó en el tema de #QueridoSergio de la categoría deportes, en esta ocasión la popularidad se debió a que existía el HT y aprovechamos para posicionar nuestra publicación para alcanzar a usuarios fuera de nuestra red.

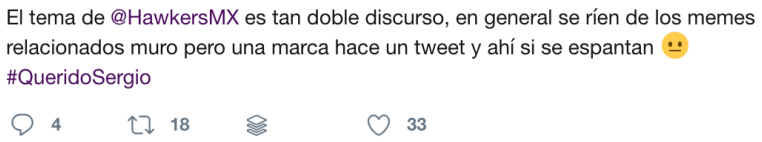


*Figura 14. Resultado de recomendación sin inclusión de atributos*

Su correspondiente es,

```
category=D wordLenght=2 ==> trending=Yes
```

Por último, una recomendación que validamos fue con en la categoría política, en este caso mostramos además de los RT el número de impresiones e interacciones con la publicación, lo que es importante resaltar es el número de personas que leyeron nuestro mensaje (Figura 15).


*Figura 15. Resultado de recomendación aplicada a la categoría Política.*

Como puede observarse la integración de las reglas convertidas en recomendaciones logran buenos resultados que consideran además de los atributos proporcionados por Twitter aquellos atributos que proponemos y fueron evaluados en la investigación. Mientras más elementos se integren existe mayor posibilidad de crecer en número de RT, siendo el atributo de **`uInfluence`** un dato relevante especialmente si alcanzamos usuarios de al menos categoría “Destacado I”, lo que

nos permite además de llegar a sus seguidores el incrementar nuestra red de usuarios.

Teniendo en cuenta que Twitter es una RS con mensajes cortos de 140 caracteres como máximo, es necesario ser conciso. Por esta razón, hacer que los tweets se vuelvan lo suficientemente interesantes para otros usuarios es un reto, y por ello tener un método de predicción que genere recomendaciones brinda una ventaja para quién lo use.

### *5.5. Desarrollo tecnológico*

Aunque la investigación no tuvo como objetivo presentar un desarrollo tecnológico, como resultado logramos crear una aplicación Web llamada *TweetPop* en fase de prototipo, adicionalmente aportamos a los usuarios una experiencia amigable sin la necesidad de realizar una configuración directamente del ambiente de trabajo o el conocer lo que es un SMA. TweetPop está constituido por un panel de control en dónde se presenta información como son: las publicaciones más populares, la hora y día recomendados para publicar y secciones para registrar el HT a posicionar.

A continuación describimos las secciones principales.

a) Página principal. La primera sección muestra una breve descripción e información más relevante acerca de *TweetPop*.
b) Inicio de sesión: Los usuarios interesados en utilizar la aplicación Web deberán registrarse por medio de un formulario que solicita un nombre de usuario, correo electrónico, contraseña, nombre y primer apellido. En caso de contar con una cuenta pueden acceder ingresando sus credenciales.
c) Panel de control. Le muestra al usuario la información necesaria para que conozca el comportamiento actual del HT a posicionar. También cuenta con acceso a los HT anteriores, cuentas vinculadas y secciones de visualización como nube de términos, tweets más populares que corresponden a su HT, día y hora más populares (Ver Figura 16).

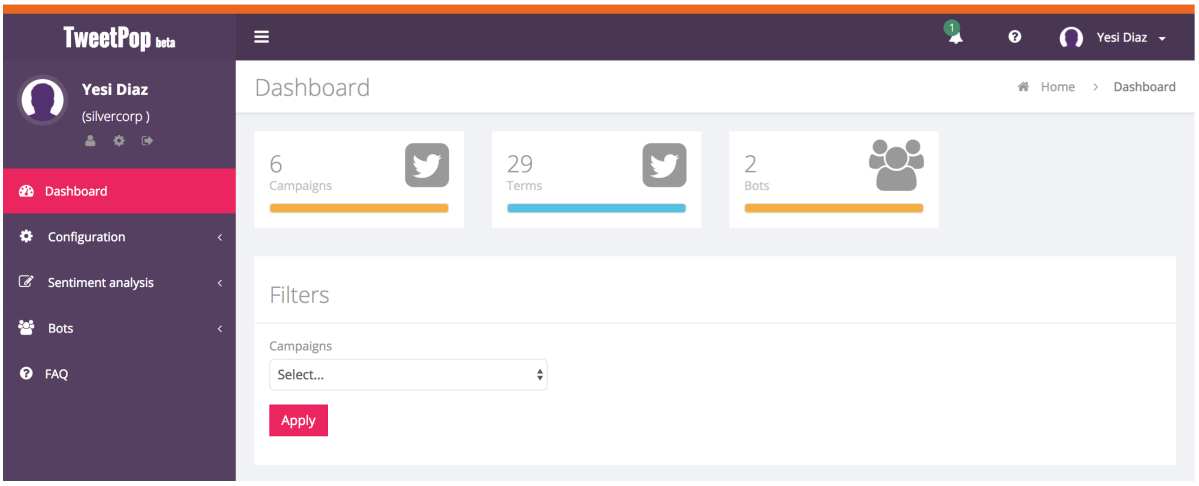


*Figura 16. Panel de Control*

d) Configuración. El usuario puede dar seguimiento a diferentes HT de manera agrupada o individual, además de definir sus parámetros correspondientes.

e) Bots. De manera adicional en dicha sección el usuario puede agregar más de una cuenta de Twitter (Figura 17) para automatizar la publicación de su contenido, además de presentarse las reglas de decisión obtenidas por el proceso realizado por los agentes (Figura 18).

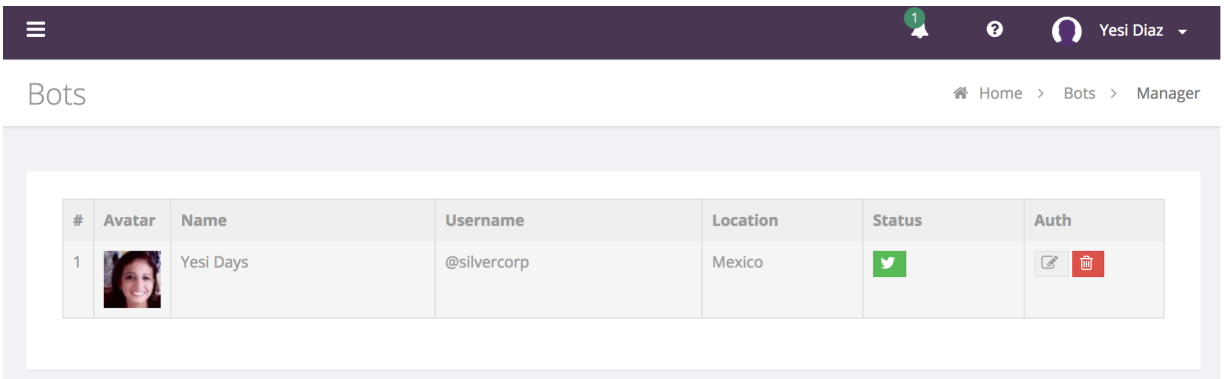


*Figura 17. Configuración de cuentas de Twitter*

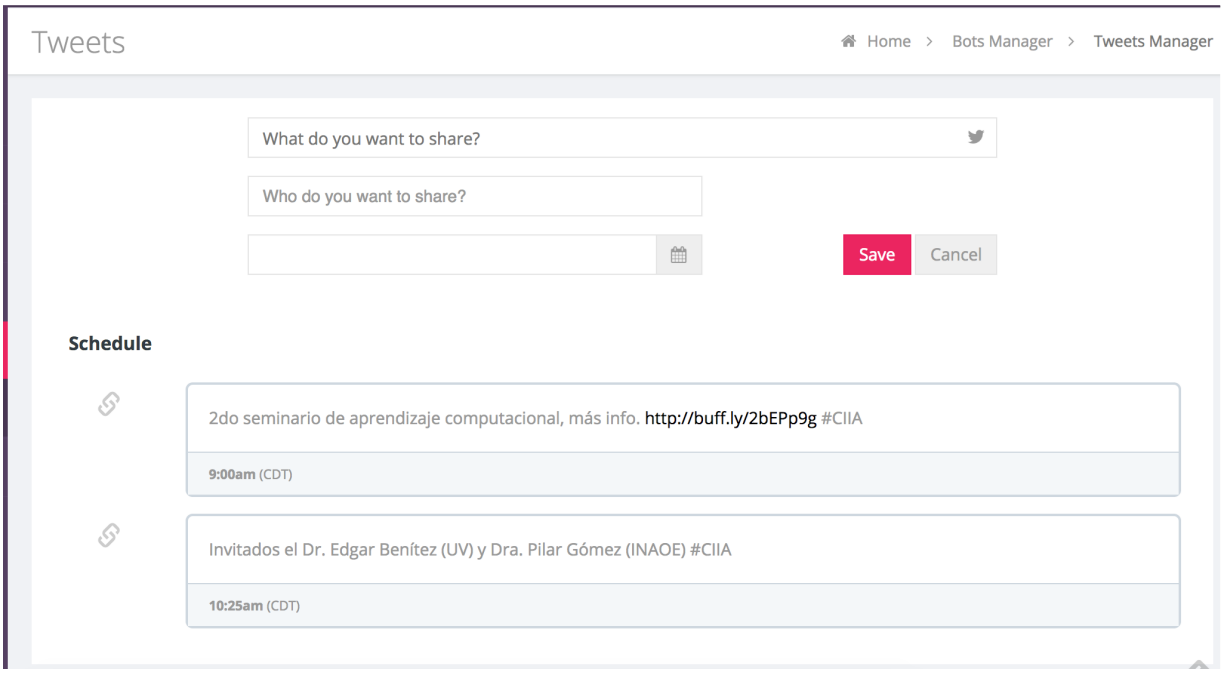


*Figura 18. Ejemplo de tweets programados para su publicación.*

Las secciones presentadas conceden al usuario las herramientas necesarias para configurar su HT a posicionar, conocer su comportamiento en el tiempo y programar los tweets, sin la necesidad de que ejecuten ninguna instrucción directamente con los agentes y artefactos, facilitando el uso de la arquitectura propuesta.

Para evaluar el diseño, viabilidad y simplicidad de la aplicación se presentó a un grupo de usuarios a los cuáles se les dio acceso y sin mayor instrucciones se les solicitó utilizaran la aplicación. Una vez que terminaron la evaluación de *TweetPop* se les pidió respondieran una encuesta en línea; algunos de ellos lo hicieron y otros únicamente se registraron e hicieron uso de la aplicación Web. Los usuarios que utilizaron la plataforma fue un grupo variado conformados por programadores, administradores de contenido, investigadores, publicistas y mercadólogos, de los cuáles el 94.1% considera útil la plataforma presentada. Aunque el objetivo principal de la investigación no es realizar una aplicación Web, esto nos permitió conocer si la metodología propuesta puede ser utilizada como una herramienta de monitoreo y posicionamiento de HT.

Uno de los objetivos clave de la implementación de una herramienta como *TweetPop* es el proporcionar una interfaz que facilite la interacción con los agentes, sin que el usuario esté familiarizado con los actos de habla, artefactos o SMA. Para ello, se utilizaron tecnologías de la Web 2.0 que permiten un diseño amigable y con la capacidad de comunicarse con la arquitectura SMA, se describen los componentes de *TweetPop* en la Figura 19.

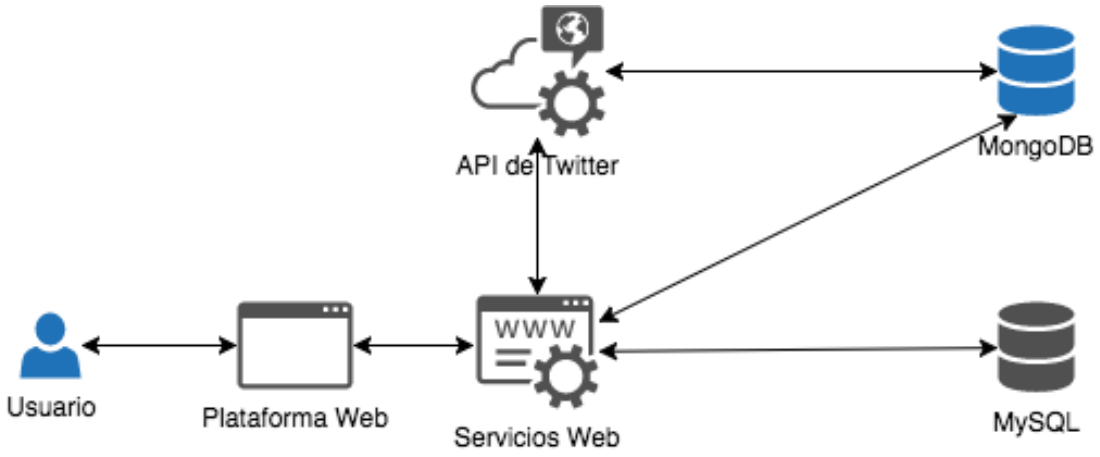


*Figura 19. Diseño general de Tweetpop*

a) La plataforma Web muestra las opciones antes descritas para que el usuario pueda interactuar indirectamente con los agentes.
b) Los servicios Web incluyen las tareas más complejas del flujo necesario para obtener las recomendaciones generadas por los agentes. Se ha diseñado una REST API que contiene un conjunto de algoritmos que permiten al usuario ajustar los parámetros necesarios para definir el proceso de análisis. Además la API se encarga de las funciones básicas de un CRUD (Create, Read, Update and Delete) para la administración de usuarios, HT y cuentas de Twitter. Además es posible realizar la extracción de los tweets, generar gráficas y presentar de los datos de una manera más simple para que sea de fácil lectura e interpretación para el usuario.
c) API de Twitter. Los servicios Web diseñados utilizan diferentes métodos de la API de Twitter, que corresponden a la extracción de los tweets, consulta,

autenticación, entre otros. Por medio de estas solicitudes y los permisos necesarios es posible obtener tweets, información de los usuarios, geolocalización y publicar en las cuentas que el usuario registre.

d) MongoDB. Es una base de datos (BD) no relacional, la cuál fue utilizada para almacenar la información de Twitter extraída mediante los servicios REST y la Api de Twitter. Se decidió utilizar éste tipo de BD debido a que cada servicio de Twitter tienen diferentes estructuras e inclusive dependiendo del tweet extraído la estructura puede cambiar, y debido a esto no es necesario mantener la misma estructura para las peticiones. Las colecciones creadas para el almacenamiento son: tweets (almacena los mensajes extraídos sin ningún proceso de limpieza), twitterUsers (guarda la información de los usuarios que publicaron) y entities (almacena los datos adicionales que contiene cada tweet).

e) MySQL. La BD almacena la información general de los usuarios registrados en la plataforma, los HT, recomendaciones, tweets a publicar e información de las cuentas de Twitter.

*TweetPop* se encuentra constituida por una parte gráfica, el motor Javascript que permite una comunicación asíncrona y los módulos en PHP que se encargan del manejo de las funciones principales antes mencionadas principalmente las relacionadas a los servicios tipo REST y la comunicación con la API de Twitter.

El diseño de la interfaz cumple con ciertos aspectos importantes como la facilidad de navegación para el usuario, presentación de la información de una manera clara y concisa y por último la adaptación de la interfaz a diferentes resoluciones. El desarrollo de la aplicación mediante Javascript permite una comunicación asíncrona que mejora la capacidad de respuesta, interacción con las diferentes secciones, menor utilización del ancho de banda al minificar el código, y un alto grado de accesibilidad.

La aplicación Web se encuentra en etapa beta, dónde su objetivo principal fue el presentarle al usuario la incorporación del SMA a un proyecto real, la presente sección fue elaborada con la finalidad de presentar los componentes de TweetPop y

los resultados de los primeros usuarios reales. Existen muchas áreas de oportunidad y comentarios de los usuarios que permite afinar los detalles de la plataforma y considerar las recomendaciones de los usuarios como son integración de un mapa de calor, comparación entre HT, guardar recomendaciones, entre otros.

# CAPÍTULO VI: Conclusiones y Trabajo a Futuro.

A lo largo de la investigación hemos descubierto que los atributos concedidos por Twitter no son suficientes para alcanzar el debido interés en las publicaciones y fue necesario profundizar en la exploración de nuevos atributos que no resultan evidentes, y son a los que nosotros denominados derivados. Este tipo de atributos son calculados a partir del número de seguidores, si tenemos contenido adicional, el tamaño de las palabras concatenadas en el HT, el tipo de contenido que se integra al mensaje y un tiempo aproximado en qué se recomienda sea publicado.

Además de los atributos hemos presentado una metodología de predicción de HT que tienen la posibilidad de convertirse en una tendencia. Para lograr que un HT adquiera mayor interés por parte de los usuarios consideramos formar recomendaciones obtenidas a través de las reglas de asociación. Sin embargo no pueden ser fortuitas, dichas recomendaciones están basadas en los árboles formados por un algoritmo de clasificación. A diferencia de otros métodos de predicción similares en nuestra investigación hemos propuesto utilizar algoritmos incrementales que se adapten al flujo de datos constante.

Propusimos el algoritmo HTree y tres variantes, los cuales examinamos y los comparamos unos con otros para así deducir a partir de los resultados del tiempo de ejecución y el porcentaje de precisión cuál de los algoritmos es el adecuado considerando nuestro caso de estudio, es decir, el flujo de datos continuo adquiridos en Twitter.

En nuestros experimentos el árbol adaptativo de Hoeffding (HTreeA) en la mayoría de los casos para las tres categorías que evaluamos tienen mejor precisión, además de mejorar considerablemente acorde el flujo de datos es mayor, es decir, cuándo tenemos al menos 7,500 instancias.

El tiempo de ejecución de HTree y HTreeA es similar, sin embargo HTreeA se mantiene sin importar la categoría o si el número de instancias no es tan extenso, lo que nos permite concluir que la manera en que está diseñado el algoritmo actúa de manera favorable conforme se tienen más instancias a evaluar. Reafirmando que el algoritmo HTreeA mantiene las estadísticas en los nodos, de manera que si un árbol previo no puede ser mejorado lo intercambia por uno más reciente.

Lo anteriormente mencionado se ve favorecido por la arquitectura que decidimos usar en nuestra investigación, el hecho de que los agentes funciones de manera colaborativa a través de una comunicación y uso de artefactos facilitó la inclusión de las tareas coordinadas por sus respectivos agentes. El emplear una arquitectura SMA es otra de nuestras aportaciones a la presente investigación.

Adicional a los resultados explicados tuvimos hallazgos interesantes durante la investigación. Los descubrimientos se enfocan principalmente al interés de las publicaciones por parte de los usuarios de Twitter, por ejemplo, dependiendo del tipo de contenido el alcance que puede tener un tweet es mayor si se agrega un enlace, ya que el usuario cuenta con información adicional que le permite reforzar la idea de quién ha publicado. Además, observamos que se tiene un nivel alto de aceptación cuando un HT está compuesto por una o dos palabras concatenadas, lo que demuestra que un HT de tamaño apropiado crea más interés para los usuarios. Con esto podemos deducir que la simplicidad en el mensaje es un elemento clave.

Uno de nuestros objetivos más ambiciosos y el cuál no logramos conseguir era el que un HT determinado por nosotros se convirtiera en tendencia, ya sea en la posición 10 o primer posición, sin embargo no alcanzamos la meta. Es por ello, que consideramos que es necesario seguir experimentando con nuevos atributos derivados que hasta el momento no se han sido propuestos, y de esta manera volver a evaluar cada algoritmo y poner a prueba las reglas de asociación que se convierten en recomendaciones para posiblemente crear tendencias a través de un HT.

Asimismo, queremos demostrar que nuestra metodología puede utilizarse para otros idiomas y sin necesidad de cambiar los parámetros, los únicos requerimientos es el definir las categorías a examinar y entrenar en una primera etapa los algoritmos. Es importante mencionar que nuestra investigación se centró exclusivamente en el idioma español lo que de manera general es más complejo que el idioma inglés, por lo que otra etapa de nuestro trabajo a futuro es obtener presentar resultados con HT en inglés y posiblemente tener mejores resultados.

Hemos presentado dos nuevas mejoras para los métodos de embolsado, utilizando árboles de adaptación Hoeffding y métodos de detección de cambios. En términos generales, observamos que al utilizar métodos de detección de deriva explícitos, mejoramos la precisión. Sin embargo, estas mejoras tienen un costo de tiempo de ejecución y memoria. Parece que el costo de mejorar la precisión en los métodos de empaquetado para flujos de datos es grande en tiempo de ejecución, pero pequeño en memoria.

Complementario a la evaluación con el idioma inglés, como trabajo a futuro queremos mejorar nuestro método de predicción, y una manera de hacerlo es utilizando muestras aleatorias con reemplaza y posteriormente combinar o ensamblar resultados, a ésta técnica se le conoce como Bagging.

Y por último, diseñar un algoritmo genético que nos ayude a mejorar nuestro método de manera que se pueda optimizar la selección de atributos hasta lograr la unión apropiada para posicionar un HT.

# REFERENCIAS

[1] O'reilly, T. (2005). What is web 2.0.

[2] Nafría, I. (2007). Web 2.0: El usuario, el nuevo rey de Internet. Gestión 2000.

[3] L. A. Adamic and E. Adar. Friends and neighbors on the web. Social Networks, 25:211–230, 2001.

[4] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In Proceedings of the 13th international conference on World Wide Web, pages 491–501. ACM, 2004.

[5] Cheng, X., Dale, C., & Liu, J. (2008, June). Statistics and social network of youtube videos. In Quality of Service, 2008. IWQoS 2008. 16th International Workshop on (pp. 229-238). IEEE.

[6] McNely, B. J. (2012, October). Shaping organizational image-power through images: Case histories of Instagram. In Professional Communication Conference (IPCC), 2012 IEEE International (pp. 1-8). IEEE.

[7] Richardson, W. (2010). Blogs, wikis, podcasts, and other powerful web tools for classrooms. Corwin Press.

[8] Yahia, S. A., Benedikt, M., Lakshmanan, L. V., & Stoyanovich, J. (2008). Efficient network aware search in collaborative tagging sites. Proceedings of the VLDB Endowment, 1(1), 710-721.

[9] Leuf, B., & Cunningham, W. (2001). The Wiki way: quick collaboration on the Web.

[10] Adamic, L., & Adar, E. (2005). How to search a social network. Social networks, 27(3), 187-203.

[11] Bernard, H.R., Killworth, P.D., McCarty, C., 1982. Index: an informant-defined experiment in social structure. Social Forces 61 (1), 99–133.

[12] Dodds, P.S., Muhamad, R., Watts, D.J., 2003. An experimental study of search in global social networks. Science 301, 827–829.

[13] Cheung, C. M., Chiu, P. Y., & Lee, M. K. (2011). Online social networks: Why do students use facebook?. Computers in Human Behavior, 27(4), 1337-1343.

[14] Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer−Mediated Communication, 13(1), 210-230.

[15] Watts, D. J. (2004). Six degrees: The science of a connected age. WW Norton & Company.

[16] Milstein, S., Lorica, B., Magoulas, R., Hochmuth, G., Chowdhury, A., & O'Reilly, T. (2008). Twitter and the micro-messaging revolution: Communication, connections, and immediacy--140 characters at a time. O'Reilly Media, Incorporated.

[17] Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009, April). Micro-blogging as online word of mouth branding. In CHI'09 Extended Abstracts on Human Factors in Computing Systems (pp. 3859-3864). ACM.

[18] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. Icwsm, 10(10-17), 30.

[19] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600). ACM.

[20] Jin, L., Chen, Y., Wang, T., Hui, P., & Vasilakos, A. V. (2013). Understanding user behavior in online social networks: A survey. IEEE Communications Magazine, 51(9), 144-150.

[21] Yang, X., Ghoting, A., Ruan, Y., & Parthasarathy, S. (2012, August). A framework for summarizing and analyzing twitter feeds. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 370-378). ACM.

[22] Zhou, D., Chen, L., & He, Y. (2015, January). An Unsupervised Framework of Exploring Events on Twitter: Filtering, Extraction and Categorization. In AAAI (pp. 2468-2475).

[23] D. M. Romero, B. Meeder, and J. M. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In WWW, 2011.

[24] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27(1):415–444, 2001.

[25] Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., & Benevenuto, F. (2011, June). Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In Proceedings of the Workshop on Languages in Social Media (pp. 58-65). Association for Computational Linguistics.

[26] Adamic, L. A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ ranking. html.

[27] Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential twitterers," in Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10), pp. 261–270, ACM, New York, NY, USA, February 2010.

[28] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In Proceedings of the 19th International Conference on World Wide Web (WWW '10) (pp. 591–600). NewYork: ACM Press.

[29] Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., & Sperling, J. (2009). Twitterstand: News in tweets. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09) (pp. 42–51). New York: ACM Press.

[30] Zubiaga, A., Spina, D., Fresno, V., & Martínez, R. (2011, October). Classifying trending topics: a typology of conversation triggers on twitter. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2461-2464). ACM.

[31] Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. Journal of the Association for Information Science and Technology, 62(5), 902-918.

[32] Recuero, R., Amaral, A., & Monteiro, C. (2012). Fandoms, trending topics and social capital in Twitter. AoIR Selected Papers of Internet Research, 2.

[33] Asur, S., Huberman, B. A., Szabo, G., & Wang, C. (2011, July). Trends in social media: Persistence and decay. In ICWSM.

[34] Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010, December). Predicting information spreading in twitter. In Workshop on computational social science and the wisdom of crowds, nips (Vol. 104, No. 45, pp. 17599-601). Citeseer.

[35] Ma, Z., Sun, A., & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in T witter. *Journal of the American Society for Information Science and Technology*, *64*(7), 1399-1410.

[36] Lin, Y. R., Margolin, D., Keegan, B., Baronchelli, A., & Lazer, D. (2013). # bigbirds never die: Understanding social dynamics of emergent hashtag. *arXiv preprint arXiv: 1303.7144*.

[37] Altshuler, Y., Pan, W., & Pentland, A. S. (2012, April). Trends prediction using social diffusion models. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 97-104). Springer, Berlin, Heidelberg.

[38] Zhang, P., Wang, X., & Li, B. (2013). On Predicting Twitter Trend: Important Factors and Models. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1427-1429).

[39] Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2).

[40] Ruggieri, S. (2002). Efficient C4. 5 [classification algorithm]. *IEEE transactions on knowledge and data engineering*, *14*(2), 438-444.

[41] Domingos, P., & Hulten, G. (2000, August). Mining high-speed data streams. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 71-80). ACM.

[42] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. Journal of the American statistical association, 58(301), 13-30.

[43] Lall, A., Sekar, V., Ogihara, M., Xu, J., & Zhang, H. (2006, June). Data streaming algorithms for estimating entropy of network traffic. In ACM SIGMETRICS Performance Evaluation Review (Vol. 34, No. 1, pp. 145-156). ACM.

[44] Minegishi, T., & Niimi, A. (2011, February). Detection of fraud use of credit card by extended VFDT. In *Internet Security (WorldCIS), 2011 World Congress on* (pp. 152-159). IEEE.

[45] Limón, X., Guerra-Hernández, A., Cruz-Ramírez, N., & Grimaldo, F. (2018). Modeling and implementing distributed data mining strategies in JaCa-DDM. *Knowledge and Information Systems*, 1-45.

[46] Dinata, I. B. P. P., & Hardian, B. (2014, October). Predicting smart home lighting behavior from sensors and user input using very fast decision tree with Kernel Density Estimation and improved Laplace correction. In *Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on* (pp. 171-175). IEEE.

[47] Guerini, M., Strapparava, C., & Özbal, G. (2011, July). Exploring Text Virality in Social Networks. In *ICWSM*.

[48] Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., & Etter, M. (2011). Good friends, bad news-affect and virality in twitter. In *Future information technology* (pp. 34-43). Springer, Berlin, Heidelberg.

[49] Masse, M. (2011). REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces. " O'Reilly Media, Inc.".

[50] Bray, T. (2017). The javascript object notation (json) data interchange format (No. RFC 8259).

[51] Zaragoza, K., Thai, N., & Christensen, T. (2011, September). An implementation for accessing Twitter across challenged networks. In *Proceedings of the 6th ACM workshop on Challenged networks* (pp. 71-72). ACM.

[52] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[53] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent dirichlet allocation. In *Advances in neural information processing systems*(pp. 601-608).

[54] Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, *70*(1), 163-173.

[55] Yecely Aridaí Díaz-Beristain, Guillermo-de-Jesús Hoyos-Rivera, and Nicandro Cruz-Ramírez, “Strategies for Growing User Popularity through Retweet: An Empirical Study,” Advances in Multimedia, vol. 2017, Article ID 4821305, 7 pages, 2017. doi: 10.1155/2017/4821305

[56] Vivoni, E. R., & Camilli, R. (2003). Real-time streaming of environmental field data. Computers & Geosciences, 29(4), 457-468.

[57] Muthukrishnan, S. (2005). Data streams: Algorithms and applications. Foundations and Trends® in Theoretical Computer Science, 1(2), 117-236.

[58] Kumar, A., Sung, M., Xu, J. J., & Wang, J. (2004, June). Data streaming algorithms for efficient and accurate estimation of flow size distribution. In ACM SIGMETRICS Performance Evaluation Review (Vol. 32, No. 1, pp. 177-188). ACM.

[59] Hulten, G., Spencer, L., & Domingos, P. (2001, August). Mining time-changing data streams. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 97-106). ACM.

[60] Gama, J., Sebastião, R., & Rodrigues, P. P. (2013). On evaluating stream learning algorithms. Machine learning, 90(3), 317-346.

[61] Bifet, A., & Gavaldà, R. (2009, August). Adaptive learning from evolving data streams. In *International Symposium on Intelligent Data Analysis* (pp. 249-260). Springer, Berlin, Heidelberg.

[62] Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (2010). Moa: Massive online analysis. *Journal of Machine Learning Research*, *11*(May), 1601-1604.

[63] Vijayarani, S., & Sathya, P. (2013, February). Mining frequent item sets over data streams using eclat algorithm. In *International Conference on Research Trends in Computer Technologies (ICRTCT-2013)*.

[64] Borgelt, C. (2003, November). Efficient implementations of apriori and eclat. In FIMI'03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations.

[65] Borgelt, C. (2005, August). An Implementation of the FP-growth Algorithm. In Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations (pp. 1-5). ACM.

[66] Agarwal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference* (pp. 487-499).

[67] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM.

[68] Ferraz, I. N., & Garcia, A. C. B. (2008). Ontology In Association Rules Pre-Processing And Post-Processing. In *IADIS European Conf. Data Mining* (pp. 87-91).

[69] Shoham, Y. (1993). Agent-oriented programming. *Artificial intelligence*, *60*(1), 51-92.

[70] Maes, P. (1995). Agents that reduce work and information overload. In *Readings in Human–Computer Interaction* (pp. 811-821).

[71] S. Russel, P. Norvig, “Artificial intelligence – A modern approach”, Prentice Hall, 1995

[72] Mouratidis, H., & Kolp, M. (2010). An architectural description language for secure Multi-Agent Systems. … and Agent Systems, 8, 99–122. http://doi.org/10.3233/WIA-2010-0182

[73] Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. *KR*, *91*, 473-484.

[74] Rao, A. S., & Georgeff, M. P. (1992). An abstract architecture for rational agents. *KR*, *92*, 439-449.

[75] A. Ricci, M. Viroli, and A. Omicini. CArtAgO: A framework for prototyping artifact-based environments in MAS. In D. Weyns, H. V. D. Parunak, and F. Michel, editors, Environments for MultiAgent Systems, volume 4389 of LNAI, pages 67–86. Springer, Feb. 2007. 3rd International Workshop

[76] R. Ricci, M. Viroli, and A. Omicini. CArtAgO: An Infrastructure for Engineering Computational Environments. In Proceedings E4MAS, pages 102119, 2006.

[77] Rao, A. S. (1996, January). AgentSpeak (L): BDI agents speak out in a logical computable language. In *European Workshop on Modelling Autonomous Agents in a Multi-Agent World* (pp. 42-55). Springer, Berlin, Heidelberg.

[78] Bordini, R. H., Hübner, J. F., & Wooldridge, M. (2007). Programming multi-agent systems in AgentSpeak using Jason (Vol. 8). John Wiley & Sons.

[79] Bordini, R. H., & Hübner, J. F. (2005, June). BDI agent programming in AgentSpeak using Jason. In Proceedings of the 6th international conference on Computational Logic in Multi-Agent Systems (pp. 143-164). Springer-Verlag.

[80] Eriksson, H. E., & Penker, M. (2000). Business modeling with UML. *New York*, 1-12.

[81] Dawid, A. P. (1992). Prequential data analysis. *Lecture Notes-Monograph Series*, 113-126.

[82] Bifet, A., de Francisci Morales, G., Read, J., Holmes, G., & Pfahringer, B. (2015, August). Efficient online evaluation of big data stream classifiers. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 59-68). ACM.

[83] Patil, A., & Attar, V. (2012). Framework for performance comparison of classifiers. In Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011 (pp. 681-689). Springer, India.

[84] Bifet Figuerol, A. C., & Gavaldà Mestre, R. (2009). Adaptive parameter-free learning from evolving data streams.

[85] Srimani, P. K., & Patil, M. M. (2015). Performance analysis of Hoeffding trees in data streams by using massive online analysis framework. *International Journal of Data Mining, Modelling and Management*, *7*(4), 293-313.

[86] Bifet, A., Holmes, G., Pfahringer, B., & Gavalda, R. (2009, November). Improving adaptive bagging methods for evolving data streams. In *Asian conference on machine learning* (pp. 23-37). Springer, Berlin, Heidelberg.

# ANEXO I: Índice de tablas

# ANEXO II: Índice de figuras

Página

# ANEXO III: Tiempo de ejecución

| Datos | HTree | HTreeO | HTreeA | HTreeR |
|---|---|---|---|---|
| DC1-250 | 15.3 | 11.3 | 16.1 | 19.77 |
| DC1-500 | 15.73 | 11.5 | 16.3 | 20.43 |
| DC1-750 | 16.73 | 12.89 | 17.4 | 26.3 |
| DC1-10000 | **17.4** | **14.5** | **18.2** | **28.3** |
| DC2-250 | 9.3 | 13.11 | 18.39 | 21.29 |
| DC2-500 | 9.53 | 13.11 | 19.3 | 22.5 |
| DC2-750 | 10.1 | 13.49 | 20.43 | 27.2 |
| DC2-10000 | **11.3** | **14.3** | **22.4** | **29.39** |
| DC3-250 | 18.3 | 25.39 | 17.2 | 30.1 |
| DC3-500 | 18.5 | 27.3 | 18.1 | 31.59 |
| DC3-750 | 19.99 | 28.39 | 18.42 | 34.5 |
| DC3-10000 | **21.39** | **32.49** | **20.99** | **35.2** |
| DC4-250 | 32.1 | 31.3 | 21.29 | 32.1 |
| DC4-500 | 35.4 | 33.1 | 24.2 | 37.2 |
| DC4-750 | 38.13 | 37.2 | 27.3 | 39.29 |
| DC4-10000 | **42.01** | **39.2** | **32.1** | **45.29** |
| DC5-250 | 22.19 | 27.7 | 38.12 | 43.1 |
| DC5-500 | 25.39 | 31.3 | 38.62 | 43.1 |
| DC5-750 | 28.1 | 33.59 | 45.29 | 48.29 |
| DC5-10000 | **28.19** | **35.2** | **47.3** | **55.3** |

Categoría: deportes

| Datos | HTree | HTreeO | HTreeA | HTreeR |
|---|---|---|---|---|
| EC1-250 | 17.3 | 20.1 | 15.3 | 21.1 |
| EC1-500 | 19.3 | 21.9 | 18.3 | 22 |
| EC1-750 | 20.1 | 22.1 | 19.2 | 22.2 |
| EC1-10000 | **21.4** | **23.1** | **20.2** | **24.2** |
| EC2-250 | 14.3 | 18.3 | 22.2 | 21.29 |
| EC2-500 | 14.5 | 26.3 | 24.3 | 22.5 |
| EC2-750 | 15 | 27.4 | 26.3 | 27.2 |
| EC2-10000 | **15.3** | **29.5** | **27.3** | **31.2** |
| EC3-250 | 18.3 | 25.39 | 22.1 | 39.1 |
| EC3-500 | 19.2 | 26.3 | 22.89 | 40.1 |
| EC3-750 | 20.2 | 29.1 | 23.1 | 45.2 |
| EC3-10000 | **21.39** | **32.49** | **25.3** | **46.3** |
| EC4-250 | 37.39 | 36.1 | 32.1 | 46.9 |
| EC4-500 | 39.1 | 36.4 | 34.6 | 45.2 |
| EC4-750 | 40.3 | 37.2 | 34.6 | 49.3 |
| EC4-10000 | **42.01** | **39.2** | **35.2** | **52.39** |
| EC5-250 | 47.3 | 39.4 | 39.2 | 47.3 |
| EC5-500 | 50.01 | 44.39 | 43.53 | 47.3 |
| EC5-750 | 51.3 | 46.2 | 44.1 | 48.1 |
| EC5-10000 | **54.1** | **47.4** | **45.2** | **49.21** |

Categoría: espectáculos

| Datos | HTree | HTreeO | HTreeA | HTreeR |
|---|---|---|---|---|
| PC1-250 | 24.1 | 21 | 28.1 | 41 |
| PC1-500 | 24.1 | 21.4 | 28.32 | 42.1 |
| PC1-750 | 25.2 | 22.1 | 29.1 | 48.2 |
| PC1-10000 | **26.1** | **23.1** | **31.2** | **49.2** |
| PC2-250 | 31.2 | 35.2 | 39 | 44.2 |
| PC2-500 | 37.1 | 41.3 | 39.2 | 47.2 |
| PC2-750 | 39.2 | 46.2 | 41.2 | 48.1 |
| PC2-10000 | **49.3** | **46.3** | **43.3** | **51.3** |
| PC3-250 | 54.2 | 45.2 | 50.1 | 39.1 |
| PC3-500 | 57.3 | 46 | 51.3 | 40.1 |
| PC3-750 | 52.3 | 46.2 | 53.5 | 45.2 |
| PC3-10000 | **59.31** | **57.2** | **53.5** | **64.3** |
| PC4-250 | 60 | 77.4 | 64.3 | 80.3 |
| PC4-500 | 60.1 | 79.1 | 67.3 | 82.1 |
| PC4-750 | 63.1 | 83.1 | 70.1 | 86.4 |
| PC4-10000 | **65.2** | **85.2** | **72.1** | **87.3** |
| PC5-250 | 50 | 75.3 | 69.4 | 87.03 |
| PC5-500 | 54.2 | 86.4 | 76.4 | 90.5 |
| PC5-750 | 59.2 | 87.5 | 83.2 | 95.3 |
| PC5-10000 | **60.4** | **90.2** | **87.2** | **96.3** |

Categoría: política

# ANEXO IV: Precisión

| Datos | HTree | HTreeO | HTreeA | HTreeR |
|---|---|---|---|---|
| DC1-250 | 32.1 | 29.1 | 31.92 | 27.31 |
| DC1-500 | 32.1 | 29.4 | 31.92 | 27.42 |
| DC1-750 | 32.71 | 29.31 | 32.01 | 28.91 |
| DC1-10000 | **33.1** | 30.1 | 33 | **29.2** |
| DC2-250 | 45.1 | 44.2 | 41.1 | 31.01 |
| DC2-500 | 45.2 | 44.23 | 41.3 | 31.1 |
| DC2-750 | 45.29 | 44.48 | 41.77 | 31.62 |
| DC2-10000 | **46.1** | 45.1 | 42.14 | **32.19** |
| DC3-250 | 68.1 | 64.1 | 67.3 | 46.2 |
| DC3-500 | 68.1 | 64.1 | 67.3 | 44.5 |
| DC3-750 | 68.49 | 64.23 | 67.81 | 46.13 |
| DC3-10000 | **70.89** | 65.2 | 68.2 | **49.4** |
| DC4-250 | 75.01 | 77.85 | 80.1 | 56.3 |
| DC4-500 | 76.23 | 78.1 | 80.1 | 56.3 |
| DC4-750 | 78.31 | 78.29 | 80.49 | 57.39 |
| DC4-10000 | 80.59 | 78.24 | **81.51** | **58.93** |
| DC5-250 | 76.01 | 77.1 | 81.53 | 53.4 |
| DC5-500 | 76.03 | 77.49 | 81.62 | 57.3 |
| DC5-750 | 77.93 | 77.49 | 82.19 | 59.23 |
| DC5-10000 | 81.39 | 78.3 | **82.92** | **60.32** |

Categoría: deportes

| Datos | HTree | HTreeO | HTreeA | HTreeR |
|---|---|---|---|---|
| EC1-250 | 33.1 | 29.1 | 31.02 | 25.3 |
| EC1-500 | 33.2 | 29.1 | 31.02 | 26.22 |
| EC1-750 | 33.13 | 29.49 | 31.45 | 25.02 |
| EC1-10000 | **34.52** | 30.42 | 32.43 | **29.2** |
| EC2-250 | 32.81 | 42.01 | 33.01 | 32.4 |
| EC2-500 | 32.87 | 42.01 | 33.52 | 33.41 |
| EC2-750 | 33.1 | 44.01 | 34.19 | 34.53 |
| EC2-10000 | 35.3 | **45.31** | **32.03** | 36.31 |
| EC3-250 | 62.01 | 57.93 | 63.01 | 4.1 |
| EC3-500 | 62.13 | 58.23 | 63.01 | 48.8 |
| EC3-750 | 63.91 | 59.33 | 63.52 | 47.32 |
| EC3-10000 | **65.03** | 62.4 | 64.2 | **49.4** |
| EC4-250 | 73.1 | 74.1 | 77.3 | 59.01 |
| EC4-500 | 73.66 | 74.1 | 77.3 | 59.72 |
| EC4-750 | 73.66 | 74.2 | 79.2 | 61.01 |
| EC4-10000 | 75.2 | 79.21 | **80.39** | **62.4** |
| EC5-250 | 75.92 | 79.1 | 78.01 | 56.2 |
| EC5-500 | 76.12 | 79.3 | 78.01 | 56.42 |
| EC5-750 | 76.4 | 79.31 | 78.1 | 57.14 |
| EC5-10000 | 77.2 | **81.29** | 79.99 | **58.32** |

Categoría: espectáculos

| Datos | HTree | HTreeO | HTreeA | HTreeR |
|---|---|---|---|---|
| PC1-250 | 23.5 | 21.34 | 24.1 | 18.42 |
| PC1-500 | 24.21 | 22.1 | 24.1 | 18.56 |
| PC1-750 | 24.58 | 22.61 | 24.51 | 20.1 |
| PC1-10000 | 24.71 | 23.51 | **25.6** | **20.42** |
| PC2-250 | 36.1 | 33.01 | 33.1 | 24.1 |
| PC2-500 | 36.1 | 33.19 | 33.41 | 24.1 |
| PC2-750 | 36.39 | 33.81 | 33.91 | 24.29 |
| PC2-10000 | **37.41** | 34.1 | 35.24 | **26.31** |
| PC3-250 | 55.2 | 52.13 | 58.1 | 45.83 |
| PC3-500 | 55.2 | 52.13 | 58.13 | 46.13 |
| PC3-750 | 55.92 | 53.81 | 58.79 | 46.13 |
| PC3-10000 | 56.21 | 54.1 | **59.14** | **48.19** |
| PC4-250 | 80.1 | 80.89 | 79.1 | 57.81 |
| PC4-500 | 80.1 | 80.89 | 79.31 | 58.1 |
| PC4-750 | 80.72 | 81.42 | 80.8 | 58.4 |
| PC4-10000 | 81.2 | **82.2** | 81.35 | **61.02** |
| PC5-250 | 81.67 | 78.1 | 82.1 | 53.4 |
| PC5-500 | 81.67 | 78.86 | 82.1 | 57.3 |
| DPC5-750 | 82.1 | 79.3 | 82.42 | 59.23 |
| PC5-10000 | 82.52 | 80.31 | **83.09** | **62.91** |

Categoría: política

# ANEXO V: Artículos aceptados, congresos y desarrollo tecnológico.

Díaz-Beristain, Y. A., & Cruz-Ramírez, N. (2016, November). Retweet Influence on User Popularity Over Time: An Empirical Study. In International Conference on Mining Intelligence and Knowledge Exploration (pp. 38-48). Springer, Cham.

Díaz-Beristain, Y. A., Hoyos-Rivera, G. D. J., & Cruz-Ramírez, N. (2017). Strategies for Growing User Popularity through Retweet: An Empirical Study. Advances in Multimedia, 2017.

Prasath, R., & Gelbukh, A. (Eds.). (2017). Mining Intelligence and Knowledge Exploration: 4th International Conference, MIKE 2016, Mexico City, Mexico, November 13-19, 2016, Revised Selected Papers (Vol. 10089). Springer.

Tweetpop. Herramienta de monitoreo y posicionamiento de tendencias. Fue utilizada y puesta a prueba por usuarios reales.