8 Conclusiones y trabajo futuro

El objetivo general de la tesis se cumplió al desarrollar una aplicación capaz de encontrar información semánticamente parecida en datos obtenidos en la red social Twitter, logrando así los objetivos particulares planteados al inicio del presente trabajo. Específicamente:

- Se implementó un subsistema para la consulta a la base de datos de Twitter, clasificación y persistencia de gorjeos que comparten rasgos seleccionados *a priori*.

- Se analizaron e implementaron métodos estadísticos para la representación vectorial de los gorjeos, las medidas de similitud y agrupamiento de gorjeos semánticamente parecidos.

- Se implementó una base de datos para la persistencia de los grupos generados por el algoritmo de agrupamiento, así como subsistemas para el control de crecimiento de la base de datos y la presentación de los mismos.

Como ya se dijo. el sistema está implementado en la plataforma Java EE, la cual proporciona un conjunto de APIs que serían difíciles de programar. También se hace uso de *frameworks* que abstraen al programador de los detalles de la tecnología Java EE, permitiendo que se concentre en la lógica del negocio. Todo esto sin la necesidad de una licencia de uso. Además, Java EE es multiplataforma porque es capaz de integrarse con otras tecnologías y con la posibilidad de desplegarse en sistemas distribuidos.

El problema principal es la curva empinada de aprendizaje para poder implementar sistemas en esta plataforma. Adicionalmente, la separación entre el modelo, la vista y el controlador es hecha por componentes con diferentes especificaciones, desde el uso de mapeo objeto relacional para la comunicación con la base de datos, pasando por los EJBs para así formar el modelo de la aplicación y poder comunicarse posteriormente con clases conocidas como Java Beans que hacen el intercambio con los componentes gráficos de JFSs; tareas casi imposibles para un principiante en la tecnología Java.
Por otra parte, el sistema de clasificación implementado en esta tesis es muy simple. Actualmente, se clasifican gorjeos por la aparición o ausencia de una palabra truncada en una lista de términos que se asocian a un tema definido *a priori* por el usuario, dejando pasar gorjeos con temas como "tráfico de órganos", "tráfico de personas" o "traficantes de drogas" en un tema correspondiente a "tráfico vehicular", ya que todos contienen el término truncado "*trafic-*". En este contexto, se plantea para un futuro la posibilidad de implementar máquinas de vector-soporte[28] para refinar la clasificación de resultados, así

[28] Conjunto de algoritmos de aprendizaje supervisado.

como la mejor selección de rasgos pensando en el uso de bigramas o el orden de ocurrencia de ciertas palabras.
Cuando el algoritmo de agrupamiento tiene malos resultados, es por esta deficiente clasificación y la no correcta elección de la distancia de corte ideal del dendograma que cambia con cada grupo analizado de gorjeos. Para arreglar esto, puede contemplarse en un futuro el cambio de algoritmo de agrupamiento, pensando en el uso de mapas auto-organizados, esto es, un modelo de red neuronal artificial no supervisado.

Como trabajo futuro se espera que el sistema sea más flexible y permita peticiones de información del usuario. El sistema hasta el momento sólo procesa y muestra los grupos generados al usuario para que el usuario los lea, sin ningún tipo de estadística e impidiéndole al usuario alguna intervención en esos resultados o incluso hacer consultas. También, a futuro se espera que el sistema pueda resolver y guardar consultas generadas por usuarios.

Actualmente los grupos generados son eliminados al pasar cierto tiempo de vida dentro de la base de datos “Temas DB”, aunque esta información eliminada tiene méritos desde el punto de vista histórico. La posibilidad de seguir un tema, evaluar qué personas lo han consultado y la posibilidad de que los usuarios puedan almacenar gorjeos y grupos de gorjeos de su interés son algunos otros de los procesos que se plantean a futuro.
Debido a la naturaleza de los mensajes de Twitter no se concretó en una evaluación sobre la precisión de los resultados, pero como trabajo futuro se pretende estudiar algún método que dé certeza de qué tan bien están formados los subtemas en un conjunto de gorjeos.

En cuanto a las virtudes del sistema desarrollado, se pueden enumerar:

Al eliminar signos de puntuación y transformar las palabras a sus raíces, el sistema unifica términos que podrían considerarse por sí mismos diferentes, esto es importante porque hay que recordar que una coma, una letra mayúscula o un sufijo hacen que una palabra sea totalmente diferente para la computadora.

Al eliminar palabras con poca información semántica y tomar como rasgos las palabras que aparecen por lo menos en dos gorjeos no sólo se reduce el espacio vectorial, también se asegura que las palabras que representan el espacio vectorial sean palabras que relacionen a gorjeos y evitar de esta forma el que dicho espacio esté compuesto por palabras exóticas de aparición unitaria en el conjunto de gorjeos.

La representación vectorial se basa en la aparición o ausencia de algún término, lo cual resulta en el cálculo más rápido del espacio vectorial en comparación de una representación más sofisticada como se

vio en el capítulo 2 donde se toma en cuenta el peso local de una palabra, el peso global de una palabra en el conjunto de gorjeos y una posterior normalización.

Se eligió derivar la matriz de distancias a partir de la matriz de energía textual, una aproximación que desde sus inicios fue concebida como una aproximación teórica para ponderar la similitud entre documentos. La energía textual no sólo da buenos resultados, su cálculo es sumamente sencillo y mucho menos costoso de lo que pudiera parecer. De no usar energía textual se tiene que pensar en una representación vectorial compleja y en una medida de distancia más robusta que la distancia euclidiana para obtener una matriz de distancias, elección que resulta particularmente mala si se considera que el sistema necesita obtener resultados sumamente rápido.
El algoritmo de agrupamiento jerárquico ofrece la ventaja de no requerir que el número de grupos sea especificado y que enfatice relaciones encontradas, formando de esta manera versiones más completas de una idea. Asimismo el algoritmo implementado tiene un *coeficiente de correlación cofenético* superior al 0.95, dando así un grado de confiabilidad a los resultados muy alto.

Los grupos generados tienen particularidades muy interesantes como lo son el que en los grupos grandes se puedan extraer términos conforme a las generalidades de un tema y que en grupos pequeños se observen sucesos particulares de un evento o tema analizado.

Por lo anterior, se puede concluir que el sistema ofrece una aproximación a la forma en que se puede extraer información de la red social Twitter, red que por su propia naturaleza es difícil de analizar.