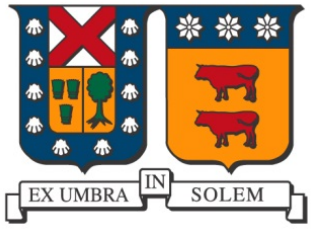
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
CIUDAD - CHILE



“DISEÑO DE UNA GUÍA DE CONDUCTAS PARA LA VIRALIZACIÓN DE ALTO IMPACTO DE CONTENIDOS EN TWITTER.”

CARLOS ALBERTO ANDRADE CABELLO

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: José Luis Martí Lara
Profesor Correferente: ???

Agosto - 2019

DEDICATORIA

Considerando lo importancia de este trabajo para los alumnos, este apartado es para que el autor entregue palabras personales para dedicar este documento. La extensión puede ser de máximo una hoja y se deben mantener este formato, tipo y tamaño de letra.

AGRADECIMIENTOS

Considerando la importancia de este trabajo para los alumnos, este apartado se podrá incluir en el caso de que el autor desee agradecer a las personas que facilitaron alguna ayuda relevante en su trabajo para la realización de este documento. La extensión puede ser de máximo una hoja y se deben mantener este formato, tipo y tamaño de letra.

RESUMEN

Resumen— El resumen y las palabras clave no deben superar la mitad de la página, donde debe precisarse brevemente: 1) lo que el autor ha hecho, 2) cómo lo hizo (sólo si es importante detallarlo), 3) los resultados principales, 4) la relevancia de los resultados. El resumen es una representación abreviada, pero comprensiva de la memoria y debe informar sobre el objetivo, la metodología y los resultados del trabajo realizado.

Palabras Clave— Cinco es el máximo de palabras clave para describir los temas tratados en la memoria, ponerlas separadas por punto y comas.

ABSTRACT

Abstract— Corresponde a la traducción al idioma inglés del Resumen anterior. Sujeto a la misma regla de extensión del Resumen.

Keywords— Corresponde a la traducción al idioma inglés de Palabras Clave anteriores.

GLOSARIO

Aquí se deben colocar las siglas mencionadas en el trabajo y su explicación, por orden alfabético. Por ejemplo:

Red Social: Sitio Web, aplicación o cualquier plataforma que utilice internet para conectar a personas a través de diversas dinámicas, como lo son el compartir y consumir contenido generado por otros usuarios, mensajería, etc.

Twitter: Red social de Microblogging en la cual los usuarios registrados comparten contenidos llamados Tweets.

Tweet: Contenido generado por lo usuarios de la plataforma Twitter, se basa en un texto limitado a 280 caracteres. Estos pueden poseer imágenes, videos, enlaces, entre otros.

Thread (Hilo): Funcionalidad de la plataforma Twitter, la cual permite escribir varios Tweets que se referenciarán de forma secuencial, permitiendo visualizar contenido extenso.

Cronología: Cuando un usuario sigue a otro, este se suscribe al contenido generado de la persona a seguir. Los tweets se presentan en la página inicial de twitter de cada usuario, a esta colección ordenada de forma cronológica se llama Cronología o timeline (TL) en inglés.

Retweet: Abreviado como RT, es la forma de compartir contenido generado por otro usuario de la plataforma, al hacer RT, los seguidores de la persona que realiza el retweet verán en su cronología el tweet original.

Favorito (Twitter): Tambien llamado Like o ”Me Gusta.es una funcionalidad presente en cada tweet, en el cual el usuario puede demostrar afinidad con el contenido presentado.

Respuesta (Twitter): Es un tweet escrito a modo de respuesta a otro tweet, generando dialogo entre los participantes.

Seguidor: Cuenta suscrita a los contenidos de otra, también es llamado Follower, por su nombre en inglés.

Seguido: Cuenta a la cual se encuentra suscrita otra cuenta.

Viralización: Fenómeno en el cual los contenidos se propagan de forma rápida e independiente, sin publicidad ni marketing, llegando de forma exponencial a nuevas personas.

Contenido Viral: Unidad de información que se ha propagado de forma rápida a una cantidad muy grande de usuarios.

Tópico: Tema (Idea o categoría) de una palabra o documento.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Debe proporcionar a un lector los antecedentes suficientes para poder contextualizar en general la situación tratada, a través de una descripción breve del área de trabajo y del tema particular abordado, siendo bueno especificar la naturaleza y alcance del problema; así como describir el tipo de propuesta de solución que se realiza, esbozar la metodología a ser empleada e introducir a la estructura del documento mismo de la memoria.

En el fondo, que el lector al leer la Introducción pueda tener una síntesis de cómo fue desarrollada la memoria, a diferencia del Resumen dónde se explicita más qué se hizo, no cómo se hizo.

CAPÍTULO 1
DEFINICIÓN DEL PROBLEMA

En la actualidad la sociedad chilena se encuentra hiperconectada, con una penetración del 71,7 % de personas con conexión a internet, de las cuales, el 94 % [1] se conecta principalmente a través de un dispositivo móvil. Estas personas utilizan en promedio 17 aplicaciones, siendo estas en su mayoría redes sociales. Lo anterior convierte a las redes sociales en uno de los mercados más interesantes para difundir contenidos debido al alcance y penetración que presentan.

Una de las redes sociales más utilizadas es la plataforma de microblogging Twitter, en la cual se pueden compartir diversos contenidos, en publicaciones llamadas tweet, la cual posee un límite de 280 caracteres (anteriormente 140). Recientemente fue estrenada una funcionalidad en twitter llamada ”Threads” o Hilos, en la cual es posible escribir diversos tweets seguidos, los cuales se verán agrupados. Esta funcionalidad es muy útil cuando la información que se desea compartir es mucho más extensa como para ser presentada en un único tweet.

Debido a esta nueva funcionalidad presentada, es posible difundir contenidos que anteriormente eran más difíciles de viralizar debido a su extensión, lo que ha generado un gran interés para diversas entidades que buscan expandir sus receptores de información (ya sea clientes, adherentes políticos u otros, según sea quien difunda el contenido).

Considerando lo interesante del mercado presentado, en conjunto con las nuevas funcionalidades introducidas en las diversas plataformas sociales, es que empresas, organizaciones, entidades de gobierno e incluso personalidades públicas están invirtiendo en personal (Community Manager) para posicionar entre los usuarios su marca o contenido asociado, los cuales actualmente puede que posea o no los conocimientos necesarios para la difusión de contenidos.

Al existir dineros invertidos, es de gran interés poseer conocimientos específicos para generar un mayor impacto y viralización de los contenidos desarrollados, por lo que se reconoce como problema el no existir una guía o manual de buenas prácticas para difundir contenido a través de twitter.

Finalmente, se reconoce que una guía de buenas prácticas para la viralización de contenidos aporta un gran valor para las diversas entidades que desean posicionar productos, marcas o campañas en la plataforma social, puesto que optimizaría los recursos invertidos para este objetivo.

[1] Chile lidera la penetración de internet en la región y el smartphone continúa siendo el favorito Fuente: Emol.com - https://www.emol.com/noticias/Tecnologia/2017/05/04/856853/Chile-lidera-la-penetracion-de-internet-en-la-region-y-el-smartphone-continua-siendo-el-favorito.html

1.1. Identificación del problema

Se identifica como problema la dificultad de popularizar contenidos en una plataforma tan utilizada y saturada de información como lo es Twitter, por lo que se busca una forma de maximizar su **impacto**, por lo que se desea analizar una gran cantidad de tweets.

Debido a que se busca realizar un análisis del contenido puesto en la plataforma, se decide realizar análisis a Hilos de tweets, buscando así una mayor precisión en el estudio de documentos.

De manera inicial se posee un dataset de 503 threads, los cuales en conjunto suman 8,894 tweets.

Cada registro posee id, número del thread perteneciente, timestamp, contenido, cantidad de retweets, cantidad de likes y cantidad de replies, tal como se puede apreciar en la figura 1.

	A	B	C	D	E	F	G
1	id	thread_number	timestamp	text	retweets	likes	replies
2	999307110902050818	Thread 1	1527088356	Extraordinary evidence at Treasury committee from Jon Thompson, CEO of HMRC on customs and Brexit today https://t.co/DJhIQhmVwJ	66	59	5
3	999307395712143360	Thread 1	1527088424	The Brexiter favourite Max Fac - would cost business between £17 and £20bn a year - that's almost 1% of GDP - jus&#65533;? &#65533; https://t.co/0MwIcwre4t	83	107	10

Figura 1: Ejemplo de registros de los datos inicialmente obtenidos.
Fuente de los datos: Kaggle.

Los registros están divididos en 5 archivos, separados por rangos de cantidad de tweets en threads, siendo estos largo de 5 a 10, 10 a 15, 15 a 20, 20 a 25 y 25 a 30 tweets.

Además de la información que ya se posee de manera inicial, es necesario rescatar información de Twitter a través de la API, atributos del usuario como lo son el id, seguidores, seguidos y cantidad de tweets, los cuales serán necesarios para el análisis a realizar.

Puesto que se busca popularizar contenidos, es de vital importancia hacer estudios sobre el texto en sí, por lo que también se deben realizar diversos análisis, como lo son la detección de tópicos y de sentimientos.

En conjunto con todo lo anterior, es necesario destacar la necesidad de almacenar toda la información obtenida de forma tal que sea posible manipular los datos de forma rápida a través de código y herramientas de análisis, por lo que se debe definir una estructura y forma de almacenamiento para los datos óptima para el estudio.

1.2. Objetivos

El objetivo general de esta memoria consiste en Diseñar y validar una guía de conductas para la viralización de alto impacto de contenidos en la plataforma Twitter a través de la funcionalidad de threads, mediante el análisis de múltiples dimensiones de diversos threads.

Objetivos específicos

- Estudiar el impacto generado según el largo, la emoción y el tópico de los threads, para identificar factores en común en los mensajes populares.
- Estudiar la relación entre el autor y el impacto de los threads, para reconocer la importancia del autor en relación al impacto del contenido.
- Analizar la relación de Favoritos, RTs y respuestas del tweet principal y los siguientes en un thread, para estudiar el comportamiento viral de estos.
- Examinar y asociar relaciones descubiertas entre variables de estudio, para proponer buenas prácticas de difusión de contenido.

1.3. Alcances

Para el trabajo a realizar se analizarán los atributos de al rededor de 9,000 tweets, agrupados en aproximadamente 500 threads. Estos documentos de estudio fueron obtenidos de la plataforma Kaggle [2]. Los documentos de estudio se encuentran escritos en inglés, algunos sólo poseen texto, sólo elementos multimedia, o ambos, por lo que se deberá realizar una limpieza inicial.

De manera inicial se poseen cinco archivos de valores separados por coma (.csv), los cuales tienen los siguientes atributos: id, número del thread perteneciente, timestamp, contenido, cantidad de retweets, cantidad de likes y cantidad de replies.

Se busca realizar un análisis tanto del impacto del contenido, como la influencia en sí del usuario en la viralización del thread, por lo que también se debe rescatar la información del usuario creador.

[2]Twitter Threads: https://www.kaggle.com/danielgrijalvas/twitter-threads/

CAPÍTULO 2
MARCO CONCEPTUAL

A continuación se presenta el marco conceptual, que sirve como base para el trabajo a desarrollar. Se reconoce la plataforma fuente de los documentos a estudiar, las técnicas de almacenamiento de la información a analizar, las dimensiones a estudiar, el análisis de estas y las metodologías de trabajo existentes.

2.1. Plataforma Twitter

Twitter es una plataforma de microblogging desarrollada por Jack Dorsey el 2006, en la cual es posible compartir contenido, ya sea texto, videos, imagenes, entre otros. Estos documentos llamados tweets presentan un límite de 280 carácteres, los cuales se ven disminuidos si es que se agregan contenidos multimedia, puesto que estos son considerados como un hipervínculo. Los usuarios utilizan la plataforma para compartir diversos tipos de contenido, como sus pensamientos a lo largo del día, noticias que encuentren interesantes o comentar ya sea programas de televisión, política, entre otros temas que consideren contingentes.

Figura 2: A la izquierda: Ejemplo de un TL, A la derecha: Ejemplo de un perfil público
Fuente: Twitter para Android.

Cada usuario genera contenido en su propia cuenta, por lo que para consumir el contenido de otra cuenta, los usuarios deben seguir a la cuenta de interés, lo que conlleva a que el contenido de esta aparezca en el inicio de Twitter del usuario que realizó el seguimiento.

Al inicio donde se muestran los contenidos se le llama cronología o Timeline (TL), un ejemplo de esto aparece en la imagen izquierda de la figura 2. En esta vista, los usuarios pueden ver los contenido generados por las cuentas que sigue el usuario.

Hay que dejar en claro que la acción de seguir un usuario es dirigida, por lo que la relación entre cuentas no es simétrica, por lo que se crea la distinción entre seguidores y Seguidos, siendo los seguidores las cuentas suscritas al contenido de un usuario, mientras que los seguidos son las cuentas a las que un usuario se suscribió.

Una cuenta puede seguir a muchos usuarios y tener muy pocos seguidores, lo que genera que éste consuma mucho contenido pero su contenido no sea consumido. El caso contrario es cuando una cuenta posee muchos seguidores pero muy pocos seguidos, a esto se le llama una cuenta popular entre usuarios. En la imagen de la derecha de la figura 2 se puede observar un ejemplo del segundo caso.

Si bien no es el objetivo general, muchos usuarios buscan tener cuentas populares, ya sea para que su contenido llegue a más personas, generar un posicionamiento en su marca o producto, o simplemente para sentirse importante entre los usuarios.

Cuando una cuenta desea compartir contenidos, es posible que un solo tweet no sea suficiente, por lo que deberá crear varios de forma seguida, teniendo esto en cuenta, Twitter lanzó una nueva funcionalidad llamada Threads, con la cual es posible redactar varios tweets de forma continua, los cuales luego serán asociados unos con otro para así poder ser mostrados de forma consecutiva, un ejemplo de esto se puede observar en la figura 3.

Figura 3: Ejemplo de Thread
Fuente: Twitter para Android.

2.2. Técnicas de almacenamiento de datos

Los datos sobre los cuales se realizarán los análisis provienen de twitter, la cual presenta su propia API para la extracción de información. Esta API posee varias limitaciones de cantidad y tamaño de consultas, haciendo necesario acotar estas consultas al mínimo, por lo que se busca no solicitar la misma información reiteradas veces, para esto, es necesario almacenar los datos que ya se hayan solicitado.

Existen diversas alternativas para almacenar información, las cuales se pueden dividir en dos categorías, estas son archivos y bases de datos.

2.2.1. Almacenamiento basado en archivos

El almacenamiento basado en archivos es una de las soluciones más simples para el almacenamiento de datos, puesto que no se debe montar ningún tipo de sistema o programa de administración, debido a esto, es común ver problemas de pequeña y mediana complejidad depender de un sistema de archivos para almacenar su información.

Algunas de las características del almacenaje basado en archivos es el enfoque menos estructurado de datos, los cuales necesariamente poseen interrelación entre archivos. El almacenaje en archivos es para problemas más generales que no necesiten una mayor optimización en las consultas y escrituras. No está diseñado para un alto consumo de los datos.

Tipos de archivos

Existen diversos formatos populares para el almacenamiento de información basada en archivos, por lo que se mencionan las 3 principales:

Comma Separated Values (CSV): Tal como se menciona en su nombre, son archivos que separan los valores o registros por coma, lo cual lo hace eficiente en espacio. Estos registros no poseen una mayor estructuración, por lo que los datos no tienen una relación entre ellos a no ser que se defina en el primer registro.

Los registros realizados en este formato son de difícil lectura para los usuarios, además de ser poco óptimo para realizar operaciones más complejas que requieran cierta lógica de negocios.

Extensible Markup Language (XML): los archivos XML están diseñados para almacenar información que posea relaciones jerárquicas, esto quiere decir, que los datos posean cierta dependencia entre los valores.

Este tipo de archivo posee una estructura basada en etiquetas de la siguiente forma:

```
<etiqueta>
    <etiqueta-2>
        valor
    </etiqueta-2>
    <etiqueta-3>
        <etiqueta-4>
            valor
        </etiqueta-4>
    </etiqueta-3>
</etiqueta>
```

Debido a que cada valor posee una etiqueta de apertura y cierre, los archivos son de mayor tamaño, pero permite una lectura simple para el usuario y una manipulación mucho más óptima para el procesamiento de datos.

Javascript Object Notation (JSON): Este tipo de archivo está basado en la estructura que Javascript maneja los objetos, lo que convierte a este tipo de archivo como uno de los más óptimos para los sistemas que dependan de este lenguaje.

```
{
    clave: valor,
    clave2: {
        clave3: valor,
        clave4: [1, a, valor, undefined]
        clave5: {
            clave6: valor
        }
    }
}
```

Además de los beneficios que trae tener un enfoque basado a javascript, los datos también tienen la opción de tener estructuras jerárquicas como ocurre en XML, con el beneficio de ser mucho más liviano.

2.2.2. Almacenamiento en bases de datos

Al contrario del almacenamiento de datos basados en archivos, las bases de datos almacenan la información de una forma mucho más estructurada, los datos pertenecen al mismo contexto y son almacenados de forma sistemática.

Otra diferencia con el almacenamiento en archivos, es que las bases de datos poseen sistemas administradores de bases de datos, los cuales realizan las consultas y manipulación de todos los datos.

Tipos de bases de datos

Dependiendo de la estructura y el manejo que se le de a los datos, pueden existir diversos tipos de bases de datos, por lo que se mencionan los tres tipos más importantes para este trabajo.

Bases de datos relacionales: Estas bases de datos son las más populares. Están basadas en registros con campos claramente definidos y rígidos, cada campo es una columna y cada registro es una fila que posee valores en cada una (Dependiendo de las reglas de negocio, es posible la existencia de algunos valores vacíos o nulos).

A nivel de definición de las bases de datos relacionales, no existe una relación directa en el orden que poseen los registros.

La gran mayoría de los sistemas de bases de datos relacionales basan sus consultas en lenguaje SQL.

Dos de las bases de datos más populares para esta categoría son ORACLE y MySQL.

Bases de datos orientada a documentos: También conocidas como bases de datos orientadas a objetos, están diseñadas para trabajar con archivos JSON y XML.

Estos sistemas de bases de datos son relativamente nuevos y buscan entregar soluciones flexibles en las cuales las bases de datos relacionales son demasiado estructuradas.

Las bases de datos orientadas a documentos poseen ciertas dificultades cuando se deben realizar operaciones en grandes conjuntos de datos.

Dos de las bases de datos más populares para esta categoría son MongoDB y CouchDB.

Bases de datos orientada a grafos: Estas bases de datos están enfocadas principalmente en la manipulación de datos que posean una gran relación entre ellos.

En estas bases de datos, los datos son representados como nodo, relación y propiedades, los que las hace muy útiles cuando las relaciones entre datos poseen tanta importancia como los datos en sí.

Dos de las bases de datos más populares para esta categoría son Neo4j y OrientBD.

2.3. Dimensiones a estudiar

El análisis inicial del problema a desarrollar posee 3 ejes claramente distinguibles, el primero es la detección de tópico o tema tratado, el segundo es la detección de emociones, y tercero, el análisis de influencia del usuario en relación a los contenidos en la red social Twitter.

2.3.1. Detección de Tópicos

La detección de tópicos en el análisis de lenguaje natural (NLP) hace referencia al estudio de documentos para reconocer ciertos ”Tópicos” o temas que se puedan estar tratando, es decir, es una forma de ”etiquetar” el tema en un documento según las palabras utilizadas.

Una de las técnicas utilizadas en la detección de tópicos es Latent Dirichlet Allocation (LDA), el cual es un modelo estadístico generativo que aprende los tópicos latentes presentes en una colección de documentos [Petrović et al., 2010].

Para el caso de estudio, la colección de documentos sería el conjunto de tweets pertenecientes a cada Thread. Es factible también considerar el thread de tweets estudiado como un solo documento, lo que hace posible realizar una detección de tópicos más precisa, esto debido a que es poco probable que un tweet sea lo suficientemente complejo como para detectar un tópico debido a que las limitaciones presentadas por la plataforma condicionan la calidad del lenguaje utilizado[Roberts et al., 2012], por lo que al aglomerar varios tweets en un documento más grande, el texto resultante posee más palabras asociadas al tema abstracto.

Existen muchas variantes de LDA para realizar detección de tópicos. En el documento [Petrović et al., 2010] se presenta una adaptación para detección de nuevos eventos, también conocido en ingles como First Story Detection. En el texto presentan un algoritmo que a través de un stream de tweets busca detectar la primera historia relacionada a un evento particular, el cual a través de hashing analiza la similitud entre documentos, para así generar un conjunto de cada tema detectado. Cada tema tiene un limite de documentos asociados, por lo que si se debe agregar uno nuevo, se elimina el más antiguo.

En este documento también generan una red de tweets por cada tema, al cual luego le analizan la velocidad de crecimiento, esto para estudiar el interés e impacto generado por el tema, para así diferenciar los temas más relevantes.

Otra variante presentada sobre LDA se encuentra en el texto [Lau et al., 2012],en el cual se busca algo similar al texto mencionado anteriormente, diferenciandose en el método de obtención y análisis de documentos. En esta variante se utiliza un vocabulario dinámico, enfocado en la clusterización y análisis de co-ocurrencia más que en la frecuencia de términos. Este vocabulario se va actualizando en porciones de tiempo definidas, lo que permite mantener actualizados los tópicos.

La forma de manejar el texto en relación al stream de datos en este texto se basa en ventanas de tiempo, lo que hace que los documentos analizados se dividan en cada ventana, esto se hace para mantener un tamaño constante de documentos analizados.

2.3.2. Detección de Emociones

La detección de emociones es parte de un área mayor llamada Affective Computing, la cual busca que los computadores sean capaces de detectar y expresar emociones humanas [Canales y Martínez-Barco, 2014].

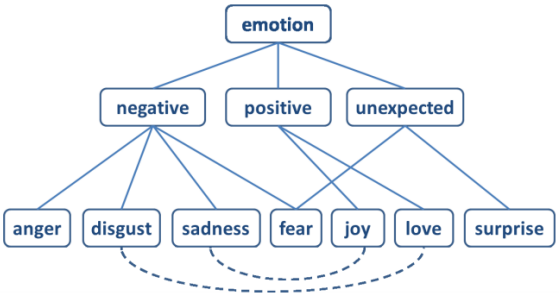


Figura 4: Relación entre las 6 emociones básicas además de amor, lineas sólidas representan herencia, mientras que las punteadas sentimientos contrarios
Fuente: Empatweet [Roberts et al., 2012].

Es posible clasificar las emociones a través de diversos modelos, poseyendo estos mayor o menor nivel de especificación en sus categorías. Dependiendo del estudio realizado podrán ser clasificadas en distintas clases. Por ejemplo, Con Sentiment Analysis es posible obtener un rango para los sentimientos entre Positivo, Negativo o Neutro, mientras que otras categorías se basan en estudios psicológicos, los cuales reconocen varias emociones básicas: Enojo, Disgusto, Miedo, Felicidad y Sorpresa [Canales y Martínez-Barco, 2014], vale decir que las emociones mencionadas pueden ser categorizadas en los setminientos también descritos, un ejemplo de esto se ve en la figura 6.

Para la detección de emociones se reconocen tres posibles métodos para realizar esto: Métodos basados en léxicos, métodos basados en máquinas de aprendizaje supervisadas y máquinas de aprendizajes no supervisadas, teniendo cada una sus ventajas y desventajas [Canales y Martínez-Barco, 2014].

Es posible construir corpus para la detección de emociones presentes en tweets a través de diversas técnicas de reconocimiento de emociones presentes en tweets a través de diversas técnicas de reconocimiento de emociones, tales como extracción y etiquetado de hashtags, detección de tópicos previamente etiquetados según emociones, máquinas de aprendizajes guiadas con tweets previamente etiquetados, entre otros [Hasan et al., 2014] [Mohammad, 2012] [Roberts et al., 2012].

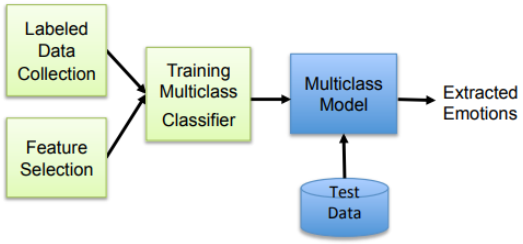


Figura 5: Ejemplo de modelo de aprendizaje supervisado para la detección de emociones en mensajes

Fuente: EmoTex [Mohammad, 2012].

Para cada tipo de modelo de detección se presentan diversos problemas [Hasan et al., 2014], ya sea el lenguaje utilizado, que le genera problemas a técnicas basadas por léxicos; la falta de etiquetas por parte del texto a analizar, que no permite entrenar máquinas de aprendizaje; e incluso la gran cantidad de temas tratados en twitter, que genera un gran número de potenciales tópicos, los que presentados en forma de vectores generaría una gran cantidad de valores cero para un tweet específico en técnicas de análisis basados en estos.

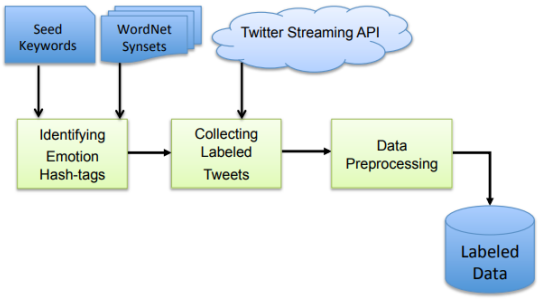


Figura 6: Ejemplo de modelo para la generación de una colección de datos etiquetados
Fuente: EmoTex [Mohammad, 2012].

2.3.3. Análisis de influencia del usuario

Además de analizar el contenido de los tweets en sí, también es necesario estudiar el impacto que genera. Actualmente existen varias formas de interactuar con el contenido en Twitter, es posible Responder, compartir en forma de Retweet (RT) y poner Me gusta a los tweets(anteriormente llamado Favorito), siendo cada una de estas acciones muy distinta en la intención con la que se hacen.

Otra dimensión que hay que tomar en cuenta al momento de analizar el impacto que tiene un tweet, es la calidad e interés que genera el mismo usuario que crea el contenido, puesto que dependiendo de el interés que genere el creador, distintas serán las reacciones que generen [Cha et al., 2010].

En la literatura se reconocen principalmente 3 tipos de influencia que pueda generar un usuario [Cha et al., 2010], indegree Influence, la cual se refiere a la cantidad de seguidores que posea el usuario; Retweet Influence, la que medida según la cantidad de RTs que reciben sus publicaciones, indica la habilidad del usuario de generar contenidos con valor de ser compartidos; y Mention Influence, la cual indica el interés generado por el usuario para que otros deseen çonversarçon el. Además de los tipos anteriores, otros autores definen una cuarta influencia que considera la frecuencia que un usuario genera nuevo tweets [Drakopoulos et al., 2017].

Para cada tipo de influencia se reconocen ciertos perfiles que generan mayor tipo de interés [Cha et al., 2010]. Para Indegree existe una gran variedad de tipos de usuarios, siendo la mayoría canales de noticias, políticos, famosos y celebridades, mientras que para Retweet existe predominancia por cuentas de noticias y otros contenidos, como por ejemplo memes y videos. Finalmente las cuentas que generan mayor interés en el ámbito de las menciones, se reconoce que en su mayoría son cuentas de celebridades.

Es posible también reconocer que tan transversal es la influencia de las cuentas populares en cada una de las 3 categorías de interés. Se reconoce en la literatura que las cuentas populares en RT y menciones poseen una correlación no despreciable, tal como se puede apreciar en la figura 7, además de que las cuentas más importantes poseen un impacto elevado en diversos tópicos, mientras que cuentas menos populares poseen popularidad en temas más acotados [Cha et al., 2010].

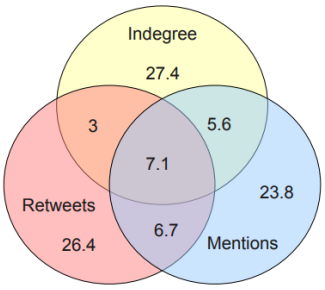


Figura 7: Diagrama de Venn de los top 100 de cada categoría de influencia, los valores se encuentran normalizados para que sumen 100 %.
Fuente: Measuring user influence in twitter. [Cha et al., 2010].

2.4. Análisis de dimensiones

Luego de completar los registros de tweets con la información faltante, además de complementarlos con la información generada a través de los análisis desarrollados, se debe estudiar los comportamientos que estas dimensiones poseen, para así detectar prácticas a repetir, con el objetivo de viralizar contenidos.

2.4.1. Métricas de análisis

Es posible definir múltiples métricas de análisis, tanto en relación a los tweets, a los usuarios y de ambos en conjunto. Tanto en [Eysenbach, 2011], [Garcia et al., 2017] y [Drakopoulos et al., 2017] son definidas múltiples métricas de gran interés para el trabajo a realizar.

Según [Drakopoulos et al., 2017] es posible definir métricas directas e indirectas. Las Métricas directas se refieren a las cuales son calculables realizando un análisis del usuario de forma aislada, esto quiere decir, considerar sólo su perfil, ignorando el grafo de usuarios de la red social, mientras que las métricas indirectas estudian diversos valores y la posición que representan en el grafo de estudio.

La mayoría de las métricas de estudio consideran valores directos como la cantidad de tweets (TW), retweets (RT), hashtags usados (HT), seguidores (FWR), seguidos (FWE), likes (FAV), respuestas (RES), menciones (MEN) y frecuencia (FR) de actividad. Mientras que algunas métricas indirectas consideran valores de centralidad de nodos calculados por diversas metodologías.

Algunas de las métricas más interesantes presentadas en [Drakopoulos et al., 2017] son la influencia conversacional, la cual considera los valores mayor cantidad de valores directos, tal como se puede observar en la ecuación 1, la actividad promedio del usuario, basada en FR y el puntaje ponderado del usuario, presentado en la ecuación 2.

$$Cv = TW + RT + MEN + FAV + RES \tag{1}$$

$$Pp = TW \cdot RT \cdot HT \cdot log(1 + FWR)^{\frac{1}{4}} \tag{2}$$

Otras métricas interesantes de destacar son la popularidad y la reputación, presentadas en [Garcia et al., 2017]. La popularidad puede ser definida tanto como la cantidad total de FWR, como también el promedio de RT $(avRT)$, mientras que la reputación se define basándose en la distribución de un grupo de usuarios según la proporción entre FWR y FWE, presentando dos métodos de obtención, el primero es a través del cálculo de incoreness, mientras que el segundo es a través del análisis de la estructura bowtie que presenta la red.

En la estructura presentada en la figura 8 se puede estudiar la distribución de todos los usuarios de twitter el 2009, En el azul se presenta el grupo más fuertemente conectado, en donde la mayoría de las cuentas se siguen entre ellas, en rojo se encuentra el grupo llamado Out, en donde las cuentas presentan una gran cantidad de conexiones hacia ellos (FWR), pero una baja conexión al grupo (FWE). En verde se presenta el grupo in el cual posee una gran cantidad de conexiones hacia el azul (FWE) y una baja conexión desde el grupo (FWR).

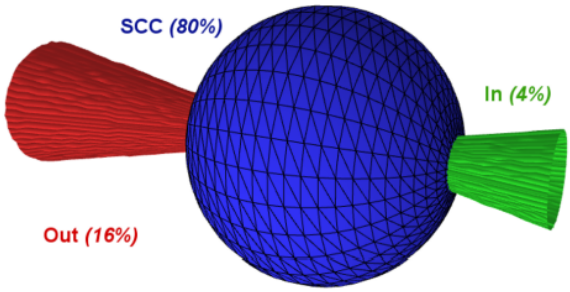


Figura 8: Representación de la distribución de la relación entre FWR y FWE de todos los usuarios en twitter el 2009
Fuente: Understanding Popularity, Reputation and Social Influence in the Twitter Society. [Garcia et al., 2017].

Un tercer conjunto de métricas destacables son las presentadas en [Eysenbach, 2011], siendo las más relevantes para este estudio Twimpact y Twindex. Los autores definen como Twimpact la cantidad de MEN recibidas por un tema por día, siendo posible definir diversos periodos de tiempo en el cual se estudiará este valor, Twindex por su parte es definido como la posición relativa del tema según la métrica de estudio, este valor se mueve entre 0 y 100.

2.4.2. Minería de datos

La minería de datos es un campo de las ciencias de la computación y estadística, la cual a través de conocimientos de inteligencia artificial, maquinas de aprendizaje, estadística y sistemas de bases de datos, se pueden descubrir patrones en grandes conjuntos de datos.

A través de la minería de datos, es posible detectar conductas como grupos (Clusters) en los datos, los cuales representan ciertas tendencias, también se pueden detectar anomalías y dependencias entre los datos.

Es posible realizar dos tipos de análisis, Descriptivo y Predictivo. El primero permite estudiar los datos en sí, evaluar la distribución y tendencias, mientras que el segundo es utilizado para modelar fenómenos, entrenar máquinas de aprendizaje automático, entre otros. El análisis predictivo es utilizado principalmente para hacer prediciones de comportamientos futuros o no conocidos.

Existen diversas técnicas de minería de datos, las cuales pueden realizar análisis descriptivos o predictivos, tales como redes neuronales, clustering, arboles de decisión, reglas de asociación, regresión lineal y modelos estadísticos.

Redes neuronales

La redes neuronales artificiales son modelos computacionales basados en el mismo funcionamiento del cerebro, poseyendo nodos o neuronas que reciben inputs y dependiendo de la configuración que poseen entregan un output. Se debe destacar que las redes neuronales en sí no son un algoritmo, sino más bien un framework para diversos algoritmos que utilizan esta lógica.

La estructura básica de las redes neuronales está constituida por grupos de nodos, los cuales son llamados capas de nodos, las cuales van procesando como información de entrada el resultado entregado en la capa anterior.

Las capas de nodos se pueden categorizar en 3, la capa de entrada, la cual recibe la información .externa"del sistema, las capas ocultas, que solo interactúan con las capas adyacentes del sistema y la capa de salida, la cual entrega el resultado del aprendizaje realizado dentro del modelo.

En estos modelos el sistema va aprendiendo y generalizando a través de ejemplos reales. La forma en lo que esto ocurre es debido a funciones de activación y de salida,en conjunto con un factor de procesamiento, llamado peso.

Existen tres tipos básicos de aprendizaje, el supervisado, en el cual se le indica a la red cual deben ser los valores de salida, el no supervisado, en el cual la maquina no recibe indicaciones de la salida, por lo cual se va adaptando y el hibrido, en el cual se le indica solo si el resultado es de buena o mala calidad.

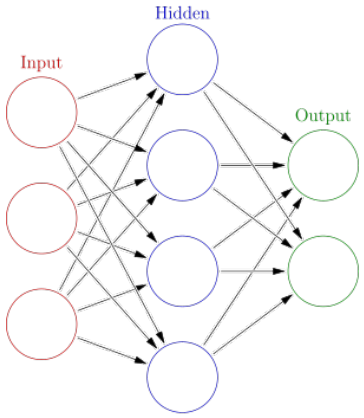


Figura 9: Ejemplo de redes neuronales
Fuente: Wikipedia.

Arboles de decisión

Los arboles de decisiones son representaciones gráficas de funciones multivariadas, son considerados modelos predictivos, puesto que construyen diagramas lógicos que permiten la toma de decisiones en base a sucesos ya medidos.

Como cualquier grafo, los arboles de decisión se componen de nodos y arcos. Cada nodo significa un estado de las variables y cada arco es la definición del estado de una variable. A cada nodo se puede llegar por sólo un camino y pueden salir tantos caminos como estados tenga la variable que está siendo decidida.

El objetivo de los arboles de decisión es poder graficar el resultado de la función analizada en relación a las conjugaciones de las variables, esto para entregar información de manera sencilla de comprender.

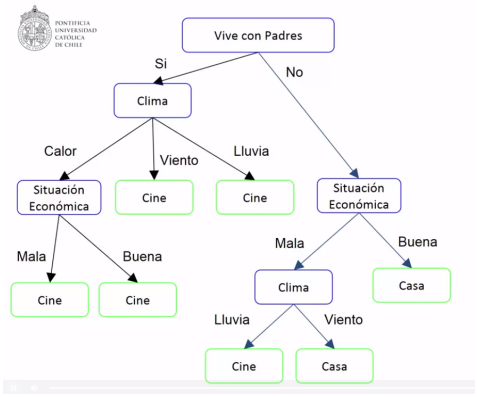


Figura 10: Ejemplo de un arbol de decisión
Fuente: Introducción a la minería de datos - Pontificia universidad católica de Chile.

Clustering

Los algoritmos de Clustering, o agrupamiento, son técnicas para aglomerar vectores según su cercanía o similitud. El valor que se utiliza puede ser calculado por diversas formas, dependiendo del caso de estudio, un ejemplo es el cálculo de la distancia euclídea.

Es posible clasificar las técnicas de clustering en jerárquicas y de particionamiento por centroides.

El clustering jerárquico crea dendodramas según la cercanía entre nodos, los algorítmos jerárquicos aglomerativos van generando uniones entre los dos clusters más cercanos, mientras que los algorítmos jerárquicos divisorios inician con un único cluster, el cual van dividiendolo según que tan lejanos son.

El clustering por partición divide los grupos de vectores en un número definido de grupos, existen diversos tipos de algorítmos para esto, los cuales van variando tanto en como se calcula la distancia, como se posiciona el centroide inicial, como se van agrupando, entre otras variaciones.

Reglas de asociación

La detección de reglas de asociación es una técnica que analiza un conjunto de datos en búsqueda de hechos en común.

Si se define un conjunto de todos los Items $I = \{i_1, i_2, i_3...i_n\}$, una transacción como $t_j = \{i_{x1}, i_{x2}...i_{xm}\}$ y una base de datos como $D = \{t_1, t_2, t_3...i_k\}$, se puede definir como una regla de asociación una combinación de dos o más items que apunta a otra combinación de uno o más items, de la forma $\{i_a, i_b\} \Rightarrow \{i_c\}$. Lo anterior significa que en varias transacciones t_x de D existen diversas combinaciones de $\{i_a, i_b, i_c\}$, de forma que las primeras dos condicionen la aparición de la tercera.

Debido a que es factible generar gran cantidad de asociaciones, es posible aplicar diversas restricciones para encontrar sólo reglas interesantes, las dos más utilizadas son la confianza y el soporte.

El soporte de un conjunto se define como la proporción en la que aparece la combinación $X = \{i_a, ..., i_b\}$ dentro de toda la la base de datos, tal como se presenta en la ecuación 3.

$$Sop(X) = \frac{|X|}{|D|} \tag{3}$$

La confianza de la regla $X \Rightarrow Y$ se define como la proporción entre el soporte del conjunto $X \cup Y$ sobre el soporte del conjunto X, tal como se presenta en la ecuación 4.

$$Conf(X) = \frac{|X \cup Y|}{|X|} \tag{4}$$

Regresión lineal

La regresión lineal es un modelo matemático que busca relacionar de forma lineal una variable dependiente Y con múltiples variables independientes Xi, de la forma $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2... + \epsilon$, donde β_i es un factor que indica la influencia de la variable independiente X_i y ϵ es un valor aleatorio.

Modelos estadísticos

Cuando se refiere a modelos estadísticos o probabilísticos en minería de datos, se refiere a la asociación de la distribución de los datos a alguna forma de distribución matemática probabilística.

Algunos de los modelos más comunes son:

- Modelo Bernoulli
- Modelo Poisson
- Modelo Geométrico
- Modelo Binomial
- Modelo Binomial Negativo

2.5. Metodologías existentes

Es necesario definir una metodología de trabajo, por lo que luego de estudiar diversos documentos, se puede indentificar una estructura de trabajo en común, la cual se presenta debido a la naturaleza de los estudios realizados. Es posible resumir la estructura común en los siguientes pasos:

- Recolección y Almacenamiento de datos.
- Manipulación de datos y obtención de métricas.
- Análisis y comparación de métricas.
- Análisis y comparación de resultados.
- Evaluación de resultados.
- Discusión y conclusiones.

Es posible realizar un bosquejo de las etapas de la metodlogía, el cual queda representado en le figura 11.

Debido a que es necesario definir una metodología de trabajo, se abordará punto por punto los pasos identifcados, exponiendo como se componen según los textos.

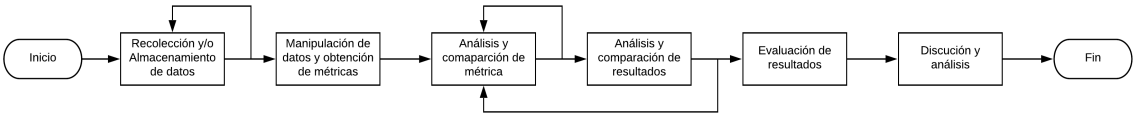


Figura 11: Estructura básica de las metodologías
Fuente: Elaboración propia.

2.5.1. Recolección y Almacenamiento de datos.

La primera etapa presente en la mayoría de las metodologías observadas es la recolección y almacenamiento de datos. En esta etapa los diversos autores definen los datos que necesitan extraer, los métodos de extracción y la forma en la que almacenarán los datos.

Algunos autores presentan esta etapa de forma recursiva, puesto que puede que de la información obtenida inicialmente debe ser complementada con nuevas consultas a la fuente de datos, según la inforamción que hayan recolectado en la iteración anterior.

De forma paralela a la obtención de datos, es necesario almacenarlos. Algunos autores mantienen sólo una base de datos con la información de los tweets y usarios que han logrado recolectar, mientras que otros dividen esta información en dos, una para la información referente a los tweets obtenidos, mientras que en la segunda está enfocada a los grafos de seguidores y seguidos que presentan las cuentas en el ambiente de estudio.

2.5.2. Manupilación de datos y obtención de métricas

Luego de completar su dataset, los autores proceden a realizar un trabajo inicial en donde manipulan los datos de forma que puedan hacer los análisis requeridos. En esta etapa construyen las métricas con las que irán a realizar estudios y comparaciones, en conjunto con todo otro procesamiento de la información que sea necesario.

Según el trabajo a realizar en cada texto, es posible que esta etapa se divida en sub categorías dependiendo de la necesidad del autor y la naturaleza del estudio realizado. Algunos ejemplos de esto se aprecian en el trabajo [Drakopoulos et al., 2017], el cual divide el trabajo en métricas directas e indirectas, mientras que en [Garcia et al., 2017] esta etapa es mínima, siendo definida todas su métricas de forma suficiente para realizar los estudios posteriores.

Este paso comúnmente es realizado una sola vez, puesto que funciona tanto como un cierre de la etapa anterior, como a la vez a forma de preámbulo para la siguiente etapa de análisis en profundidad.

2.5.3. Análisis y comparación de métricas

Luego de poseer todos los datos y valores necesarios, el siguiente paso es el estudio y análisis de la información recabada. En esta etapa es donde se presenta la mayor diferencia entre los diversos autores.

Por una parte, es posible realizar un estudio para cada métrica definida de forma aislada, disponiendo de un capítulo para cada métrica definida en el punto anterior. Las técnicas de análisis más comunes en este tipo de trabajos son la elaboración de gráficos y regresiones matemáticas y estadísticas, con el fin de poder modelar el comportamiento de la métrica de estudio.

Por otra parte, algunos autores realizan análisis varias métricas en conjunto, buscando una interrelación entre los valores definidos. En este tipo de trabajo comúnmente se busca modelar uno o más fenómenos estudiados que presenten diversas combinaciones de variables y métricas. En este tipo de trabajos es posible encontrar diversos modelos para un mismo fenómeno, por lo que también realizan diversos análisis de correlación y divergencia entre modelos.

Es posible que en esta etapa se definan métricas nuevas, por lo que se deben realizar los análisis correspondientes para esta, por lo que se considera que esta etapa puede llegar a ser recursiva.

2.5.4. Análisis y comparación de resultados

En esta etapa se estudian los resultados obtenidos de los modelamientos realizados. Si fueron realizados varios modelos para el mismo fenómeno, se define cual obtuvo mejores resultados, mientras que si se realizó un único análisis, es posible definir que puntos pueden mejorarse para así obtener un mejor resultado.

Dependiendo de las conclusiones obtenidas en esta etapa se proceda a realizar nuevos análisis, por lo que desde esta etapa se proceda a una etapa anterior.

2.5.5. Evaluación de resultados

Luego de finalizar los estudios y análisis respectivos, es necesario comprobar los resultados obtenidos. Si el trabajo culminó en un modelo, en esta etapa se realizan pruebas para determinar que tan correcto y preciso es, mientras que si se definieron más de un modelo para el mismo fenómeno, se puede indicar en que condiciones cada modelo es recomendado, o si un modelo es superior en todos los sentidos.

2.5.6. Discusión y conclusiones

Es posible que esta etapa se presente como una sola o subdividida en discusión y además conclusiones, pero el foco principal en ambos casos es el cierre del trabajo realizado, indicar el por qué de los posibles problemas en los resultados, en conjunto con los modelos obtenidos y las conclusiones tanto los modelos como el trabajo en si.

CAPÍTULO 3
PROPUESTA DE SOLUCIÓN

3.1. Metodología de trabajo

Definir metodología a utilizar. Pasos, etapas y repeticiones.

3.1.1. Almacenamiento de datos

Luego de obtener obtener los datos a través de las APIs, se posee un tweet con los siguientes atributos (id, thread, timestamp, contenido, rt, likes, replies, id usuario,usuario es verificado?,cantidad de seguidores usuario, cantidad de seguidos usuario, cantidad de tweets usuario).

Además, luego del procesamiento del contenido de los tweets a través de análisis de tópico y sentimiento, se obtiene nueva información del tweet (id, emociones detectadas, tópicos detectados).

3.1.2. Detección de tópicos

Técnica a usar.

3.1.3. Reconocimiento de emociones

Técnica a usar.

3.1.4. Influencia del usuario

Metodología a usar.

3.1.5. Popularidad

Fórmula para cuantificar.

3.1.6. Análisis de datos

Técnica de minería de datos.

3.2. Análisis de tópico

En este punto se desarrollará el estudio de tópicos según Hilos, se estudiarán los tópicos que más aparecen, el largo del hilo por tópico, etc.

Relación entre tópico y popularidad

Se estudiará la distribución de los tópicos y la popularidad

3.3. Análisis de Emociones

En este punto se analizarán las emociones por hilo, la distribución de la emociones según el largo, etc.

Relación entre emociones y popularidad

Se estudiará la distribución de la popularidad de las emociones.

3.4. Análisis de Influencia

Acá se estudiará la relación directa entre el perfil del usuario y la popularidad del Hilo.

Relación entre la influencia y la popularidad

3.5. Relación entre dimensiones

En esta sección se estudiarán las relaciones de todas las combinaciones de los 3 análisis, luego se estudiará la relación entre las dos dimensiones más concluyentes

3.5.1. Relación entre tópico y emoción

3.5.2. Relación entre tópico e influencia

3.5.3. Relación entre influencia y emoción

3.6. Análisis de relaciones obtenidas

Hacer comentarios de los resultados obtenidos

3.7. Selección y construcción de recomendaciones

Indicar cuales son las relaciones que mas impactan, indicar que funciona y que no.

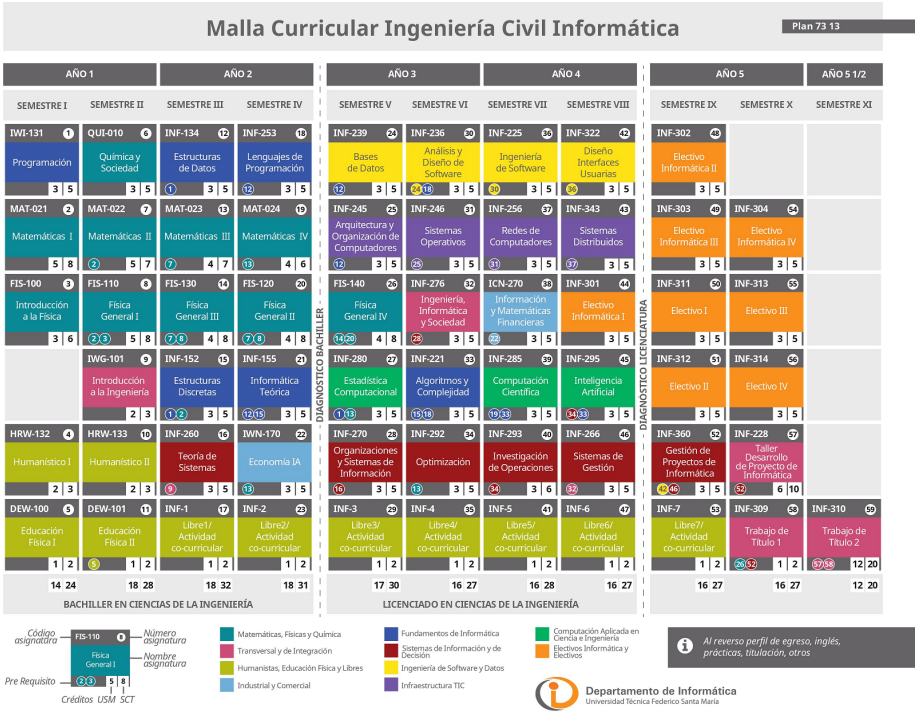


Figura 12: Malla Curricular Ingeniería Civil Informática.
Fuente: Departamento de Informática.

CAPÍTULO 4
VALIDACIÓN DE LA SOLUCIÓN

La validación se realizará analizando otro set de datos y se evaluará si las observaciones realizadas en el punto anterior se cumplen

4.1. Datos

Datasets a utilizar

4.2. Pruebas

Análisis de los datos

Tabla 1: Coloquios del Ciclo de Charlas Informática.
Fuente: Elaboración Propia.

Título Coloquio	Presentador, País
"Sensible, invisible, sometimes tolerant, heterogeneous, decentralized and interoperable... and we still need to assure its quality..."'	Guilherme Horta Travassos, Brasil.
"Dispersed Multiphase Flow Modeling: From Environmental to Industrial Applications"'	Orlando Ayala, EE.UU.
"Líneas de Producto Software Dinámicas para Sistemas atentos el Contexto"'	Rafael Capilla, España.
...	...

CAPÍTULO 5
CONCLUSIONES

Las Conclusiones son, según algunos especialistas, el aspecto principal de una memoria, ya que reflejan el aprendizaje final del autor del documento. En ellas se tiende a considerar los alcances y limitaciones de la propuesta de solución, establecer de forma simple y directa los resultados, discutir respecto a la validez de los objetivos formulados, identificar las principales contribuciones y aplicaciones del trabajo realizado, así como su impacto o aporte a la organización o a los actores involucrados. Otro aspecto que tiende a incluirse son recomendaciones para quienes se sientan motivados por el tema y deseen profundizarlo, o lineamientos de una futura ampliación del trabajo.

Todo esto debe sintetizarse en al menos 5 páginas.

ANEXOS

En los Anexos se incluye todo aquel material complementario que no es parte del contenido de los capítulos de la memoria, pero que permiten a un lector contar con un contenido adjunto relacionado con el tema.

REFERENCIAS BIBLIOGRÁFICAS

[Canales y Martínez-Barco, 2014] Canales, L. y Martínez-Barco, P. (2014). Emotion detection from text: A survey. En Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC), pp. 37–43.

[Cha et al., 2010] Cha, Meeyoung and Haddadi, Hamed and Benevenuto, Fabricio and Gummadi, P Krishna and others (2010). Measuring user influence in twitter: The million follower fallacy. Icwsm, 10(10-17):30.

[Drakopoulos et al., 2017] Drakopoulos, G., Kanavos, A., Mylonas, P., y Sioutas, S. (2017). Defining and evaluating twitter influence metrics: a higher-order approach in neo4j. Social Network Analysis and Mining, 7(1):52.

[Eysenbach, 2011] Eysenbach, G. (2011). Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. J Med Internet Res, 13(4):e123.

[Garcia et al., 2017] Garcia, D., Mavrodiev, P., Casati, D., y Schweitzer, F. (2017). Understanding popularity, reputation, and social influence in the twitter society. Policy & Internet, 9(3):343–364.

[Hasan et al., 2014] Hasan, M., Rundensteiner, E., y Agu, E. (2014). Emotex: Detecting emotions in twitter messages.

[Lau et al., 2012] Lau, J. H., Collier, N., y Baldwin, T. (2012). On-line trend analysis with topic models:\# twitter trends detection topic model online. Proceedings of COLING 2012, pp. 1519–1534.

[Mohammad, 2012] Mohammad, S. M. (2012). # emotional tweets. En Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 246–255. Association for Computational Linguistics.

[Petrović et al., 2010] Petrović, S., Osborne, M., y Lavrenko, V. (2010). Streaming first story detection with application to twitter. En Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics, pp. 181–189. Association for Computational Linguistics.

[Roberts et al., 2012] Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., y Harabagiu, S. M. (2012). Empatweet: Annotating and detecting emotions on twitter. En LREC, volumen 12, pp. 3806–3813. Citeseer.