UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FíSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

COMPARACIÓN DE ALGORITMOS PARA RESÚMENES AUTOMÁTICOS EN TWITTER

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL EN COMPUTACIÓN

LUIS ALBERTO MARTÍNEZ MARTÍNEZ

PROFESOR GUÍA:
BARBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:
JORGE PEREZ ROJAS
JOCELYN SIMMONDS WAGEMANN
CLAUDIO ALVAREZ GOMEZ

SANTIAGO DE CHILE
2019

Resumen

Las redes sociales han producido un cambio en la forma como las personas se comunican, interactúan y obtienen información. Ya que, sin ser un medio que reemplace los medios tradicionales, las redes sociales logran situarse en el corto plazo en una posición privilegiada por las nuevas generaciones. Este posicionamiento se puede explicar, en parte, porque es un medio que permite la producción y consumo de información en forma fácil, descentralizada y a un bajo costo.

Este cambio en la generación de contenido, junto con los grandes volúmenes de información que son producidos, tiene como consecuencia que la cantidad de información a la cual un usuario puede acceder es mucho mayor que la que es capaz de leer. Lo anterior produce que para un usuario común sea muy complejo mantenerse informado sobre un evento noticioso.

Siguiendo esta motivación, los principales objetivos de este trabajo son:

- Reconocer cuáles son los principales métodos de generación de resúmenes automáticos para noticias en redes sociales.
- Realizar un análisis comparativo exploratorio de los métodos estudiados.

Para realizar el análisis comparativo se estudiaron diferentes propiedades de los resúmenes, este análisis consideró un cotejo entre lo publicado en redes sociales y lo informado en medios establecidos. Adicionalmente, se estudió el grado de representación (cobertura) del resumen generado en relación con el evento completo.

Las principales contribuciones de este trabajo son: dar a conocer los métodos existentes, agrupándolos en tres categorías; implementar y probar el comportamiento de distintos métodos utilizando casos reales; con los resúmenes generados, establecer cuales son los factores más influyentes en la calidad del resumen.

La metodología propuesta para este trabajo consistió en utilizar un conjunto de mensajes de Twitter que hablen sobre un evento en particular. Luego, sobre cada eventos aplicar algoritmos para la generación de resúmenes en redes sociales. Finalmente, evaluar los resúmenes considerando diversidad y cobertura

Con los datos obtenidos a partir del estudio realizado, se concluye que un factor que afecta la calidad del resumen obtenido es la representación vectorial del texto, ya que, al utilizar distintas representaciones, varia los resultados en la evaluación, siendo *FastText* la que obtiene mejores resultados. A su vez, se observa que no es posible utilizar los medios tradicionales de noticias como *ground truth*, para evaluar resúmenes de redes sociales, esto debido a que difieren en el enfoque que se le da al evento. En general, al comparar los valores obtenidos por las métricas utilizadas se observa que los algoritmos estudiados tuvieron desempeños similares, pero siendo *phrase reinforcement* el algoritmo con mejor desempeño, obteniendo un desempeño 10 % superior en la evaluación de cobertura con respecto a los otros algoritmos.

A mi familia, gracias por todo.

Tabla de Contenido

Índice de Tablas

*

Índice de Ilustraciones

*

Capítulo 1

Introducción

1.1. Situación

La disposición de la información a un clic de distancia, produce que cualquier hecho que suceda en el mundo, se encuentre publicado de forma casi inmediata en las redes sociales. Este hecho hace que los usuarios de dichas redes, utilicen las mismas para informase de los eventos que suceden a diario se informen y además puedan opinar respecto a dicho eventos. En Estados Unidos se determino que durante el 2018 el 68 % de los adultos se informaron de eventos noticiosos a través de las redes sociales[1].

Para ser parte de las redes sociales existen una serie de aplicaciones, dentro de las cuales encontramos las de Mircroblogging, donde los usuarios publican y comparten mensajes breves sobre algún tema de su interés. Dentro de estos sitios de microblogging se tiene a Twitter, sitio lanzado en el año 2006, siendo en la actualidad el más popular (sobre el cual se focaliza este trabajo), alcanzando los 326 millones de usuarios activos mensualmente[2].

Los usuarios de estas aplicaciones publican información de distintos ámbitos, tanto personal como global, pudiendo realizar esto en más de una oportunidad. Así mismo, se observa que, en el caso de Twitter los usuarios comparten información casi instantáneamente a medida que suceden los eventos. Incluso entidades involucradas directamente en los hechos pueden compartir información. Un ejemplo del impacto que generan los eventos en las redes sociales, fueron las elecciones presidenciales de Estados Unidos del 2016, para este acontecimiento se publicaron mil millones de mensajes relacionados con este tema[3], demostrando el alcance mundial de las redes sociales. Aunque, para un evento se publican muchos mensajes importantes, también se publican muchos mensajes que pueden ser considerados ruido o spam, es decir, mensajes que no tienen información relevante del evento.

Por lo explicado anteriormente, surge la necesidad de agrupar y presentar la información al usuario de un modo que sea sencillo de entender y rápido de consultar. Una forma de realizar

[1] http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
[2] https://twitter.com/TwitterIR/status/1055419090305613825
[3] https://blog.twitter.com/2016/how-election2016-was-tweeted-so-far

esta presentación es generando un resumen automático del texto publicado en los diferentes mensajes. Pero las técnicas convencionales para resumir documentos de forma automática no son aplicables de forma directa sobre los mensajes de Twitter por las siguientes razones:

- Los mensajes de Twitter (tweets) son muy cortos, máximo 280 caracteres [4], esto produce que los usuarios deformen su lenguaje al usar abreviaciones, hashtags y emoticones.
- Cada usuario posee un estilo de escritura distinto y no estandarizado, a diferencia de, por ejemplo, un artículo periodístico que si tiene una estructura clara.

Lo anterior hace que intentar resumir eventos sólo utilizando tweets sea una tarea desafiante, demandante de tiempo y esfuerzo, para obtener información que posiblemente sea el input de un proceso mayor.

Dado lo mencionado en los párrafos anteriores, la generación de resúmenes con información precisa y de utilidad ha sido un tema en desarrollo desde el 2010 con los trabajos de O'Connor et al. [2010] y Chakrabarti and Punera [2011]. A partir de estos trabajos, se reconoce que existe una fuente potente de información, que solo se debe mejorar la forma de acceso y presentación para que sea fácilmente usada por quienes presenten el interés en ello.

Al investigar sobre métodos existentes para generar resúmenes, se detecta un espacio respecto a que no existe una evaluación comparativa entre ellos. Lo anterior se debe principalmente a la dificultad de encontrar criterios comunes entre las distintas evaluaciones realizados por los trabajos y la dificultad para reproducir las evaluaciones realizadas. Además de la inexistencia de un conjunto de datos estándar para realizar las evaluaciones.

1.2. Hipótesis de Investigación y Preguntas de investigación

Este trabajo se construye sobre la siguiente hipótesis:

1. La información publicada en redes sociales sobre un evento noticioso, permite describir el evento en su globalidad.

Este trabajo tiene como objetivo abordar las siguientes preguntas de investigación:

1. ¿Qué métricas son utilizadas para evaluar resúmenes automáticos?¿son siempre las mismas?
2. ¿Es similar la información noticiosa compartida en redes sociales con respecto a la publicada en medios de periodísticos tradicionales?
3. ¿Qué tipos de técnicas existen para realizar resúmenes automáticos de eventos noticiosos desde las redes sociales?

[4]En septiembre de 2017, la longitud de los mensajes fue duplicada. La gran mayoría de los mensajes utilizados por este proyecto son anteriores a esta fecha

4. ¿Qué tipos de técnicas y representaciones de los datos son las más efectivas?

1.3. Objetivos

1.3.1. Objetivo General

El objetivo principal de este trabajo es investigar, probar y comparar qué métodos de la literatura existentes pueden ser aplicados a datos de redes sociales para la generación de resúmenes automáticos de eventos noticiosos. Además, se propone realizar un análisis exploratorio de carácter comparativo entre los diferentes métodos.

1.3.2. Objetivos Específicos

1. Investigar y clasificar los métodos existentes para llevar a cabo resúmenes asociados a un evento noticioso en redes sociales.

2. Investigar qué criterios se utilizan en la literatura para evaluar la calidad de un resumen y comparar dichos criterios.

3. Mediante el estudio de casos comparar, utilizando distintos indicadores de calidad, resúmenes generados con distintos métodos y representaciones.

1.4. Contribuciones

Las principales contribuciones de este trabajo son:

- Generar una clasificación para literatura asociada a la generación de resúmenes de eventos noticiosos a partir de redes sociales.

- Detallar las métricas existentes para evaluar resúmenes automáticos.

- Realizar un estudio de casos para comparar la calidad de distintos métodos de generación de resúmenes.

1.5. Resultados

Una vez realizados los experimentos sobre los métodos seleccionados. Se obtuvo como principales resultados:

- Existen varios trabajos relacionados con la generación de resúmenes automáticos en redes sociales, pero las evaluaciones realizadas por cada uno son difícilmente reproducibles

y comparables entre si.

- Los medios tradicionales de noticias no pueden ser utilizados para comparar la información presente en las redes sociales, ya que no presentan intersecciones en su vocabulario.
- El factor más influyente en la generación de un resumen es el modelamiento que se use para representar la información del texto presente en los mensajes.

1.6. Metodología

Al no encontrar una metodología formal que permitiera guiar las actividades del trabajo se definió una propia que permitiera identificar distintos métodos y luego realizar comparaciones entre ellos. Esta metodología consta de cinco pasos. Primero, se categorizaron los trabajos existentes según las técnicas comunes utilizadas. Segundo, se identificaron y recolectaron mensajes asociados a distintos eventos del mundo real. Tercero, a partir de la categorización realizada en el paso uno se implementaron distintos métodos. Cuarto, se escogen un conjunto de eventos que formaran los casos de estudio y se generan resúmenes utilizando los métodos implementados en la parte anterior. Quinto, realizar una evaluación de la cobertura y diversidad de casa método. En la figura 1.1 se muestran los pasos asociados a esta metodología.

Figura 1.1: Diagrama sobre la metodología utilizada

1.7. Esquema de la Tesis

El trabajo de Tesis ha sido organizado en 7 capítulos, los cuales tienen directa relación con las etapas que conforman la metodología presentado en el punto anterior: El capítulo uno, recién presentado corresponde a la introducción del trabajo. En el capítulo dos se explica el marco teórico en el cual se desarrolló el trabajo, presentando las distintas técnicas utilizadas por los trabajos del área. En el capítulo tres, se realiza una categorización del estado del arte existente, clasificando los trabajos según las técnicas en común que utilicen. El capítulo cuatro se exponen las distintas técnicas de evaluación de resúmenes automáticos, analizando sus ventajas y desventajas. El capítulo cinco detalla los distintos experimentos realizados, los conjuntos de datos utilizados y métodos seleccionados. En el capítulo seis se analizan y discuten los distintos resultados en cada experimento para cada uno de los métodos implementados. Finalmente, el capítulo siete presenta las conclusiones obtenidas en el trabajo y trabajo futuro para profundizar en la investigación.

Capítulo 2

Marco Teórico

Esta investigación se basa en la aplicación práctica de técnicas de *Minería de datos*, *Recuperación de la Información* y *Machine Learning* Adedoyin-Olowe et al. [2013]. Estas diferentes técnicas son usadas para extraer y obtener, desde Twitter, resúmenes asociados a eventos noticiosos, los que incluyen información ruidosa y no estructura.

Para poder generar un resumen a partir de un evento tratado en una red social, en general, se utilizan los siguientes pasos:

Recolección de Eventos Consiste en extraer mensajes desde la red social, que de alguna forma estén relacionados con un evento noticioso del mundo real.

Filtrado Dada la naturaleza heterogénea de los tweets, se debe realizar una limpieza de estos, eliminando aquellos que no aporten información.

Detección de subeventos Busca agrupar los mensajes que hablen sobre temas similares, para identificar todos los subeventos existentes en el conjunto. Es importante identificar la mayor cantidad de subeventos, para lograr abarcar todos los aspectos del evento. Los subeventos corresponden a los hitos más importantes de un evento. Por ejemplo, un gol en un partido o la cantidad de muertos en un accidente.

Selección de representantes Por cada subevento detectado se debe escoger uno o más representantes, de acuerdo a algún criterio, como temporalidad, o relevancia, para que formen el resumen final.

Durante la aplicación de este proceso, una actividad de importancia es modelar y dar formato al contenido textual, no olvidemos que en el origen cada actor escribe uno o más comentarios a un evento en el estilo que le acomode en ese instante. Es en este punto donde se aplican técnicas de *Minería de datos* y *Machine Learning* asociadas a procesamiento de texto y clustering. Las técnicas explicadas en este capítulo fueron utilizadas para el desarrollo de este trabajo o se encuentran presentes en la literatura relacionada a la generación de resúmenes.

2.1. Técnicas de procesamiento de texto

Para poder extraer información presente en un texto, primero se debe modelar correctamente. Esto consiste en eliminar el ruido y palabras irrelevantes o mal escritas con el fin de determinar cuales términos son más descriptivos con respecto al significado global de texto. La forma más común de realizar esto es asignarle un valor de importancia a cada termino presente en el texto, esto se conoce como representación vectorial del texto. Existen distintas formas de generar la representación vectorial de un texto. En esta sección se presentan los métodos que fueron utilizados para este trabajo.

2.1.1. Stemming

Muchas veces dentro de un mismo texto se encuentra el mismo verbo, pero conjugado de distinta forma o se encuentra el mismo término en su versión en plural y singular, si bien son palabras distintas, sus significados son casi idénticos. Es en estos casos cuando los términos pueden ser representados mediante una raíz común, por ejemplo, las palabras "argue", "argued", "argues", "arguing" pueden ser representado por la raíz común "argu", el proceso por el cual se determina esta raíz es denominado *stemming*, propuesto por primera vez por Earl [1966]. Existen distintos algoritmos de stemming como *Porter Stemming*, *Dawson Stemming* y *Lovins Stemming*, en este trabajo se utilizó Porter Stemming, propuesto por Porter [1980]. Este algoritmo se aplicó como preprocesamiento de la información, antes de aplicar cualquier técnica de generación de resúmenes. La ventaja de este algoritmo es que permite reducir el ruido presente en el conjunto de mensajes, esto porque agrupa palabras escritas de forma similar dentro de un mismo termino y así disminuye la cantidad total de términos, permitiendo reducir la redundancia de términos.

2.1.2. TF-IDF

Este método fue propuesto por primera vez por Sparck Jones [1972], es la abreviación para *term frecuency-inverse document frecuency*, es una estadística que indica que tan importante es una palabra en un documento que pertenece a un conjunto más grande de documentos. Esta medida se basa en la idea que si un término aparece varias veces en un documento debe ser importante, pero además considera que si este término aparece en muchos documentos entonces no aporta información descriptiva sobre el documento inicial. El valor tf-idf para un término t es *tf-idf(t) = tf(t)· idf(t)*, donde

Term Frecuency (*tf(t)*) Su valor corresponde a la cantidad de veces que aparece un término en un documento dividido por el largo del documento.

Inverse Document Frecuency (*idf(t)*) Está representado por la siguiente ecuación.

$$\mathrm{id}f(t) = \log \frac{1 + n_{\mathrm{d}}}{1 + df(\mathrm{d}, t)} + 1 \tag{2.1}$$

En este caso n_{d} representa el número total de documentos, *df(d,t)* corresponde al número de documentos que contienen el termino t.

Esta métrica ha sido muy utilizada para modelar la información presente en todo tipo de documentos. Pero como indica el trabajo realizado por O'Connor et al. [2010], *tf-idf* no fun-

ciona correctamente si se aplica directamente sobre los mensajes extraídos de redes sociales, es decir, que los vectores obtenidos no representan correctamente la importancia de los términos, tendiendo a darle una importancia menor. Esto porque los mensajes publicados en redes sociales como Twitter son muy cortos y numerosos, lo cual produce que el termino *tf* sea muy bajo. Una forma de solucionar este problema es no considerar cada tweet como un documento individual, sino considerar una agrupación de estos como un documento. La agrupación puede ser bajo distintos criterios, por ejemplo, en el trabajo realizado por Bhaskar et al. [2012] consideran como un solo documento todos los tweets asociados con una query de un usuario, para posteriormente extraer las frases más relevantes y generar un resumen.

2.1.3. FastText

Es un algoritmo basado en redes neuronales propuesto por Joulin et al. [2016] y por Bojanowski et al. [2016] que genera de forma eficiente representaciones vectoriales para palabras, considerando su contexto. Este algoritmo considera que cada palabra está conformada por un conjunto de n-gramas, donde n va desde 1 hasta el largo de la palabra. Por ejemplo la palabra "sunny" es descompuesta en los siguiente n-gramas "[sun, sunn, sunny, unny, nny]", a diferencia de otros modelos en los cuales las palabras son consideradas como una unidad atómica. Luego, para cada uno de los n-gramas se genera una representación vectorial utilizando *continuous skipgram model*, propuesto por Mikolov et al. [2013]. Este modelo busca generar representaciones vectoriales de palabras considerando el contexto que las rodea, es decir, dos palabras tendrán representaciones similares si el contexto que las rodea es similar, ejemplo las palabras *hielo* y *agua* tendrán contextos similares, en cambio *hielo* y *mesa* tendrán contextos distintos. Se define el contexto de una palabra p como el conjunto de palabras que están a una distancia n de p. Este modelo toma como entrada un conjunto de palabras que constituyen el vocabulario sobre el cual se aprenderán las representaciones y como observaciones recibe los contextos de cada una de las palabras del vocabulario. Luego este modelo utiliza los contextos entregados para generar un vector por cada palabra p del vocabulario, en el cual cada entrada del vector corresponde a la probabilidad que una palabra d aparezca en el contexto de p. Este vector corresponde a la representación vectorial de p. Se utiliza el modelo antes descrito para aprender una representación por cada n-grama de la palabra, posteriormente se utiliza la suma de las representaciones de cada n-grama como la representación final de una palabra, esto tiene como beneficio que permite aprender representaciones de palabras extrañas o infrecuentes, ya que estas también pueden ser descompuestas en n-gramas de términos ya conocidos. Esto es muy práctico para texto de redes sociales donde se suelen encontrar palabras mal escritas. Además, como queda demostrado en Joulin et al. [2016] el tiempo de entrenamiento de este modelo es mucho menor que el de otras representaciones. En este caso los autores recomiendan entrenar el modelo bajo una gran cantidad palabras, ellos utilizan entre 1 millón y 2 millones de palabras, por lo cual este modelo debe ser entrenado bajo el conjunto total de tweets.

2.1.4. Glove Vectors

Es un algoritmo de aprendizaje no supervisado, es decir, que no requiere ningún tipo de etiquetado sobre los datos, para aprender representaciones vectoriales de palabras, propuesto por Pennington et al. [2014]. Este método se basa en determinar que tan asociada están dos palabras en un contexto determinado, para esto definen un ratio de co-ocurrencia, esto es la

probabilidad que una palabra w aparezca junto con una palabra k. Por ejemplo el ratio de la palabra solido y hielo es más alto que el ratio de la palabra *solido* y *vapor*. Para determinar estos ratios, se calcula una matriz de co-ocurrencias en donde cada entrada X_{ij} representa la cantidad de veces que una palabra j aparece en el mismo contexto que la palabra i, para este caso se define que j está en el contexto de i. Una vez construida la matriz, se aplican un algoritmo de factorización de la matriz que consiste en asignar distintos pesos a las palabras según su probabilidad y su distancia entre ellas. Al igual que Fasttext descrito en la página anterior, este modelo también debe ser entrenado bajo una gran cantidad de palabras.

2.2. Clustering

Corresponden a un conjunto de algoritmos que buscan agrupar elementos en conjuntos, denominados clúster, de tal forma que todos los elementos asignados al mismo conjunto son más similares, bajo algún criterio definido anteriormente, a los elementos de los otros conjuntos. Los conjuntos generados tendrán distintas propiedades dependiendo de su propósito y algoritmo escogido. En el caso de la generación de resúmenes automáticos, las técnicas de clustering son utilizadas para agrupar los mensajes que hablan de temas similares y detectar los subeventos existentes.

2.2.1. K-Means

K-means es un algoritmo que genera una partición de n elementos en k conjuntos, donde k es un parámetro, en donde cada elemento corresponde al clúster cuyo valor medio es el más cercano. Este algoritmo fue propuesto por MacQueen et al. [1967]. El funcionamiento de K-means es detallado a continuación:

1. Se escogen al azar k elementos que serán los primeros centroides de cada clúster.

2. Cada dato es asignado al grupo o clúster con la media más cercana. Más formalmente si c_{i} es un centroide en el conjunto C, entonces cada dato x es asignado a un clúster en base a lo siguiente:
$$\min_{c_{\mathrm{i}} \in C} dist(c_{\mathrm{i}}, x)^2 \tag{2.2}$$
Cada elemento es asignado al clúster con el cual dicho elemento tenga la menor distancia con el centroide del clúster. En este caso *dist* corresponde a la distancia euclidiana.

3. Se actualizan los centroides, el nuevo centroide es calculado como el promedio de todos los datos asignados a clúster de ese centroide.
$$c_{\mathrm{i}} = \frac{1}{|S_{\mathrm{i}}|} \sum_{x_{\mathrm{i}} \in S_{\mathrm{i}}} x_{\mathrm{i}} \tag{2.3}$$
En donde S_{i} es el conjunto de elementos asignados al clúster i.

4. Se repiten los pasos anteriores hasta que los puntos dejen de cambiar de clúster, o la suma de las distancias es minimizada o se alcanza un número máximo de iteraciones.

Bottou and Bengio [1995] demostró que K-means converge a un resultado, este resultado puede ser un óptimo local o global, dependiendo de la configuración inicial de clusters, por lo

cual una asignación distinta de los centroides iniciales puede producir un mejor resultado, para solucionar esto, se recomienda ejecutar el algoritmo con distintas configuraciones iniciales.

Una método derivado de K-means corresponde a *bisecting k-means*, presentado por Steinbach et al. [2000]. Este método consiste en partir con todos los elementos en un solo clúster, para posteriormente aplicar K-means sobre este conjunto para encontrar 2 clústers. Luego, se escoje uno de los dos clústers generados y se repite el proceso. El algoritmo termina cuando se han alcanzado los k clústers deseados. Para este algoritmo es importante definir que clúster escoger para particionar, se pueden utilizar distintos criterios, como el tamaño del clúster (escoger el más grande), o el clúster que genere el menor aumento del error cuadrático.

2.2.2. Jerárquico

El clustering jerárquico es otro tipo de técnicas de clustering, en la cual se busca construir una jerarquía de clústers. A diferencia de k-means en el clustering jerárquico se puede utilizar cualquier medida de distancia. De acuerdo con lo propuesto por Tan et al. [2007] existen dos enfoques distintos para aplicar clustering jerarquico:

Aglomerativo: Cada dato parte en un clúster individual y en cada paso se unen los pares de clústers más cercanos, según una noción especifica de proximidad. Requiere definir la noción de proximidad entre clúster. Existen distintos métodos de clustering aglomerativo, según el tipo de proximidad que utilicen. Los cuatro métodos principales son:

1. Single: Es el más simple. En este método la distancia entre clústers corresponde a la distancia entre los dos puntos más cercanos de cada clúster, fue propuesto por Sneath et al. [1973].

2. Complete: Similar al método Single, pero ahora se escogen los dos puntos más lejanos, propuesto por King [1967].

3. Average: Utiliza la distancia promedio entre todos los elementos de ambos clústers.

4. Ward: En cada paso se busca unir el par de clústers, tal que, después de ser mezclados produzcan el mínimo aumento de la varianza entre clústers, propuesto por Ward Jr [1963].

Divisivo: En este método todos los elementos son asignados a un único clúster. Luego, en cada paso se divide el clúster de mayor tamaño. Este proceso es repetido hasta que cada elemento esta presente en un solo clúster. Para dividir los clusters se utiliza un algoritmo de clustering normal, como es el caso de bisecting k-means que utiliza el algoritmo K-means para dividir el cluster.

2.2.3. Clustering incremental

Para ambos métodos descritos anteriormente es necesario fijar de forma previa el número de clústers a generar. Esto no siempre es factible de realizar, ya que muchas veces no se conocen las características del iniciales del conjunto de datos. En estos casos se puede aplicar

clustering incremental, el cual consiste en:

1. Se toma un elemento del conjunto que no tenga un clúster asignado. Al principio ningún elemento tiene asignado un clúster.

2. Se calcula una medida de distancia entre este nuevo elemento y todos los clústers ya existentes. Qué medida de distancia se va a utilizar depende del domino sobre el cual se esté trabajando. En el caso inicial cuando no existe ningún clúster, se crea un cluster con ese primer elemento.

3. Si la similitud entre el nuevo elemento y todos los demás clústers es menor a cierto umbral predefinido entonces el elemento pasa a formar un nuevo clúster.

4. Si la medida de similitud es mayor que el umbral entonces ese elemento pasa a formar parte de ese clúster.

5. Se repite el proceso hasta que todos los elementos han sido asignados.

Una ventaja de este método de clustering es que puede ser aplicado de forma dinámica, es decir, los elementos pueden ir apareciendo en tiempo real y ser asignados a cluster. Este tipo de técnica es aplicada en el trabajo de Sankaranarayanan et al. [2009], en cual se busca generar un resumen en tiempo real sobre noticias de último minuto. Para esto se van agregando los mensajes a medida que van siendo publicados. Este tipo de algoritmos requiere definir una medida de distancia entre el elemento y los clústers. la medida de distancia puede decir la distancia máxima entre el nuevo elemento y los elementos del clúster, o la mínima u otra que sea definida según el contexto. Se requiere, además, definir un umbral para decidir cuando el nuevo elemento pertenece o no a un clúster.

2.2.4. Clustering sobre grafos

Corresponde a agrupar los nodos presentes en un grafo en clústers, considerando los arcos existentes en el grafo, tal que, los elementos de un clúster tienen muy pocos arcos en común con los elementos de otro clúster, pero si tienen muchos arcos entre los elementos de un mismo clúster.

Este tipo de clustering es muy útil en estructuras como las existentes entre las redes sociales, ya que en las redes sociales existen relaciones y conexiones tanto entre usuarios, entre mensajes y entre mensajes y usuarios.

Un ejemplo de algoritmo de clustering sobre grafos es SCAN, acrónimo para *Structural Clustering Algorithm for Networks*, propuesto por Xu et al. [2007]. Este algoritmo está diseñado para encontrar agrupamientos ocultos en redes de elementos (grafos). La idea principal de este algoritmo es encontrar clústeres entre los elementos del grafo, además de identificar nodos que sean outliers y nodos que sean *hubs*, se define como hub un nodo que está conectado con más de un clúster y un outlier es un elemento que está conectado a un solo clúster. Por ejemplo, en la figura 2.1, el elemento 13 correspondería a un outlier y el elemento 6 correspondería a un hub. Para generar los clústers este algoritmo se basa en que, si dos vértices comparten muchos vecinos, es decir, si están conectados a los mismos vértices, entonces esos

vértices pertenecen a un mismo clúster, es decir, si la cantidad de vecinos en común entre dos vértices es mayor a un umbral especifico, estos son asignados al mismo clúster

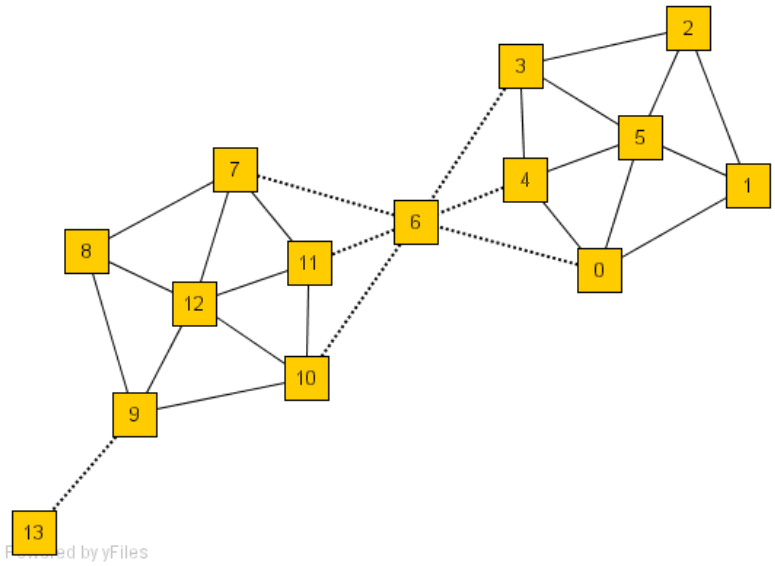

Figura 2.1: Ejemplo de un grafo

2.3. Twitter

Twitter es una red social fundada en el año 2006. Los usuarios de twitter basan su interacción en compartir mensajes entre sus contactos, estos mensajes son denominados "tweets", con la característica especial que los mensajes no pueden superar los 280 caracteres [1]. Existen distintos tipos de relaciones entre los mensajes, y relaciones entre los usuarios de la red social. En el caso de los mensajes un mensaje puede ser un mensaje original, es decir, creado por un usuario; también puede ser una respuesta a otro mensaje; o puede ser un retweet, que corresponde cuando un usuario comparte un mensaje escrito por otro usuario. En el caso de los usuarios, la relación es "seguidores", los seguidores de un usuario pueden ver los mensajes que este publica, pero él no necesariamente puede ver los mensajes de sus seguidores. Los mensajes que el usuario ve son mostrados en orden cronológico en una interfaz llamada *Timeline*, en la figura 2.2 se muestra un ejemplo de un *Timeline* para un usuario cualquiera. Además, los usuarios pueden agregar marcas en los mensajes. La primera marca son los *hashtags* que corresponden a términos que un usuario quiera destacar y compartir con otros usuarios. Los *hashtags* definen a los denominados *trending topics*, estos corresponden a todos los mensajes que contienen un determinado hastag, definiendo un tema en específico que está siendo muy comentado por los usuarios. La segunda marca corresponde a las menciones, un usuario puede etiquetar a otro usuario en el texto de su mensaje.

[1] Hasta antes de septiembre del 2017 el largo máximo de los mensajes era 140 caracteres. La gran mayoría de los mensajes utilizados en esté trabajo son previos a esa fecha.

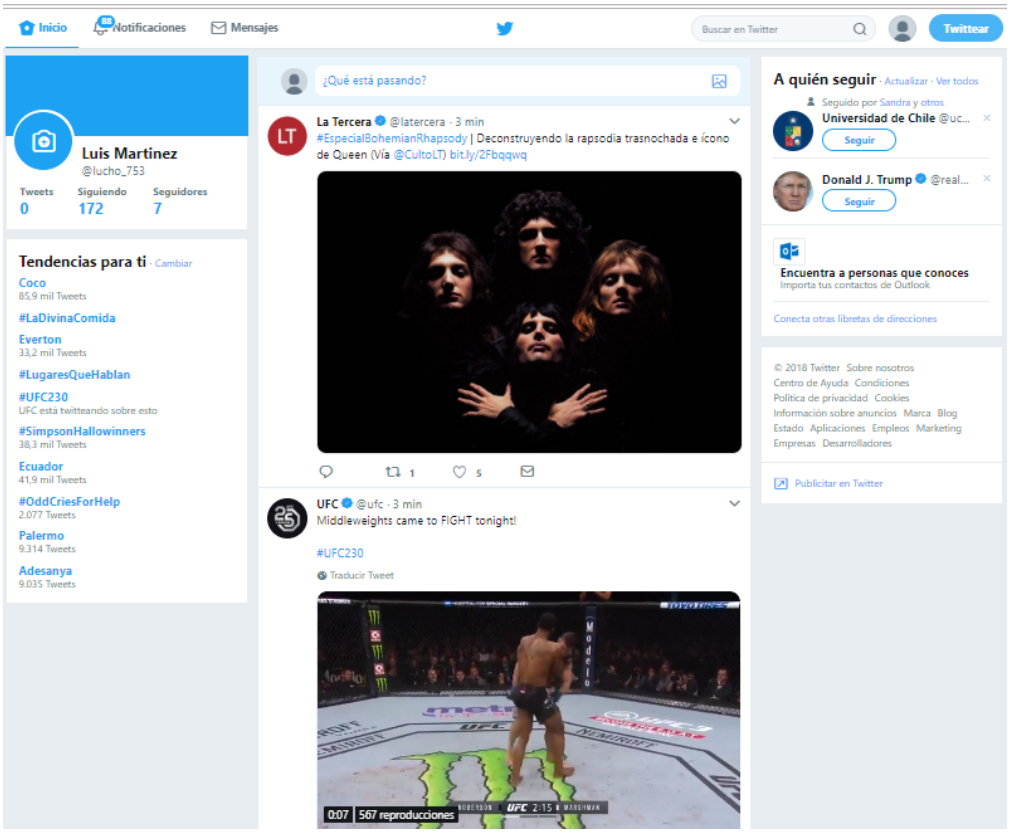

Figura 2.2: Ejemplo de timeline para un usuario

2.4. Resumen

Las técnicas y algoritmos descritos en este capítulo son las bases para la generación de resúmenes en redes sociales, cada una de estas técnicas puedes ser incluida en distintas etapas. Por ejemplo, los algoritmos de stemming son utilizados en la etapa de filtrado, ya que permiten agrupar términos similares disminuyendo el ruido. Por otra parte, los algoritmos para generar representaciones vectoriales son el conjunto de entrada que utilizan los métodos de clustering para detectar los subeventos y agrupar los mensajes que contengan información adicional.

Capítulo 3

Categorización para resúmenes automáticos

En este capítulo se presentan distintas categorizaciones existentes para agrupar los trabajos asociados a la generación automática de resúmenes. Además de explicar las categorizaciones ya existentes se propone una nueva categorización propia de este trabajo, específica para resúmenes que utilicen redes sociales de microblogging como fuente de información. Una vez explicada la categorización, también se mencionan los trabajos pertenecientes a cada categoría y se detallan los métodos que fueron usados como baselines para los experimentos posteriores que se explican en el capítulo cinco.

Según lo propuesto por Jones et al. [1999] el proceso de producir resúmenes automáticos busca: dado un texto o un conjunto de textos, producir una reducción del documento(s) original ya sea por selección o por generalización de lo que se considere importante para un usuario o para una tarea específica.

A partir de lo expuesto se tienen dos categorías básicas. La primera es resúmenes que son producidos por la generalización del contenido, es decir, el resumen no se genera a partir de los documentos originales, sino que se busca generar un texto nuevo que describa al original, este tipo de resúmenes son denominados abstractivos. La segunda categoría corresponde a resúmenes que son generados a partir del documento original, se extraen oraciones desde el documento y dichas oraciones conforman el nuevo resumen, los resúmenes de esta categoría son denominados extractivos. Un ejemplo de resumen extractivo es el propuesto por Luhn [1958], en el cual se extraen frases desde el texto según la frecuencia con la que aparecen. Ambas categorías fueron propuestas por Edmundson [1969].

Otra categorización es la propuesta por Lloret and Palomar [2012], en la que se puede distinguir entre resúmenes *generales* o *enfocados en el usuario*, el primero buscar ser un reemplazo del documento completo, ya que cubre todos los aspectos relevantes del texto. En cambio, el segundo está enfocado cubrir una necesidad de información del usuario, ya sea sobre una consulta en específico o un tópico en particular. También se pueden generar resúmenes con respecto a un solo documento o múltiples documentos. Puede considerar uno o múltiples idiomas.

La generación automática de resúmenes para texto ha sido un tema de investigación durante más de 50 años. El primer trabajo en esta área corresponde al realizado por Luhn [1958] en el año 1958, para producir resúmenes de artículos científicos, en este trabajo se extraen las frases con mayor frecuencia en un documento para formar el resumen.

Estos dos grupos de categorizaciones presentadas no son excluyentes, un método para generar resúmenes puede ser abstractivo y enfocado en el usuario o general y extractivo. En particular para este trabajo nos enfocaremos en métodos que generen resúmenes extractivos y generales, en un solo idioma. Se adopta este enfoque, porque los mensajes de Twitter son muy cortos y ruidoso como para utilizar métodos abstractivos, además como en Twitter se comparte grandes volúmenes de información, sobre eventos noticiosos, se busca resumir el evento completamente.

En particular para el desarrollo de este trabajo usaremos la definición de evento propuesta por Kalyanam et al. [2016a], que es:

Definición 3.1 *Un evento es una ocurrencia en el mundo real y con un periodo de tiempo asociado T_e y con un conjunto de mensajes ordenados por tiempo que discuten esta ocurrencia y que fueron publicados durante T_e.*

3.1. Resúmenes en redes sociales

Como se explicó en el capítulo 2 la generación de resúmenes puede ser visto como un proceso de 4 pasos, *recolección de eventos*, *filtrado*, *detección de subeventos* y *selección de representantes*. En este trabajo nos enfocaremos en las 3 últimas etapas, se asumirá que los eventos ya fueron detectados e identificados, esto porque el objetivo del trabajo es comparar resúmenes y no detectar eventos en redes sociales.

Debido a la gran popularidad alcanzada por las redes sociales y los altos volúmenes de información que son publicados en tiempo real se han realizado un gran número de trabajos asociados a la generación de resúmenes Chakrabarti and Punera [2011], Sharifi et al. [2010], Yajuan et al. [2012], Bian et al. [2015], Wang et al. [2015], Xu et al. [2013], Schinas et al. [2016], He et al. [2017], Long et al. [2011], Inouye and Kalita [2011], Metzler et al. [2012]. Si bien los trabajos recién mencionados pueden ser agrupados utilizando las categorizaciones anteriores, la gran mayoría de los trabajos quedan agrupado en una sola categoría, resúmenes extractivos y resúmenes generales. Lo anterior también produce que determinar que métodos pueden ser utilizados para realizar las comparaciones propuestas sea más complejo, ya que la gran mayoría de los métodos estará en una sola categoría. Con el fin de determinar que métodos pueden ser utilizados para comparar, se revisaron los trabajos del estado del arte considerando distintos criterios, los cuales son listados a continuación:

1. Fuente de datos utilizada, debe ser una red social, idealmente Twitter.

2. Disponibilidad de los datos utilizados.

3. Tipos de datos utilizados, corresponde a un domino abierto o cerrado.

4. Disponibilidad del código fuente del proyecto.

5. Cuales son los criterios utilizados para detectar subeventos.

6. Qué criterios se utilizan para seleccionar los mensajes.

Para determinar que trabajos podían ser usados se armó una lista de trabajos potencialmente importantes. La lista se armó a partir de trabajos encontrados en Google Scholar, cuyo año de publicación fuese mayor al 2007. En el proceso de búsqueda se utilizaron como palabras clave "automatic summarization", "Twitter summarization" y "social media summarization". Solo se consideraron trabajos que tuvieran como fuente de datos Twitter, esto por los grandes volúmenes de información generados y porque Twitter provee una API, la cual permite descargar los mensajes publicados de forma automática. Considerando lo antes descrito se armó una lista de trabajos potencialmente importantes, esta lista incluye 35 trabajos relacionados (ver anexo). Además, se consideraron trabajos que tuvieran su código fuente disponible y que los datos usados fueran de dominio abierto, es decir, generasen resúmenes sobre cualquier tipo de evento.

Una vez revisada la lista de trabajos potencialmente importantes, se observo que la metodologías utilizadas tendían a repetirse entre ellos. Por lo cual, se propone una categorización en base a las técnicas utilizadas para agrupar los mensajes. Los trabajos revisados fueron agrupados en tres grandes grupos, los cuales son explicados a continuación.

1. **Métodos basados en grafos:** Corresponden a los trabajos que utilizan las distintas relaciones entre mensajes y usuarios, para modelar una estructura de grafo, que representa las relaciones de dependencia entre mensajes y/o usuarios.

2. **Métodos basados en clustering:** Corresponden a los trabajos que aplican distintos métodos de clustering para agrupar los mensajes que hablen sobre temas similares y así detectar los subeventos. Luego bajo una combinación de criterios, definidos por cada trabajo, se escogen los representantes de cada clúster.

3. **Métodos basados en modelos probabilísticos** corresponde a aquellos trabajos que buscan determinar la probabilidad con la cual un mensaje puede pertenecer a un tópico. Luego agrupan todos los mensajes asociados a un tópico y seleccionan un representante según una combinación de criterios, definidos por cada trabajo.

Debido a lo costoso que seria evaluar los 35 trabajos mencionados, se seleccionaron por cada categoría los trabajos que cumplían la mayor cantidad de criterios, este subconjunto de trabajos fueron considerados como relevantes y fueron considerados como posibles candidatos para el análisis. En la tabla 3.1, se muestran los trabajos que fueron catalogados como relevantes.

Trabajo	Categoría	Importancia
Sharifi et al. [2010]	Grafos	Dominio de datos abierto; Utilizado como baseline; Simple de implementar
Schinas et al. [2016]	Grafos	Código fuente disponible; Dominio de datos abierto; considera distintos aspectos para evaluar; considera contenido multimedia; optimiza distintos criterios para seleccionar los mensajes.
Xu et al. [2016]	Grafos	Optimiza distintos criterios para seleccionar mensajes; domino de datos abierto.
Inouye and Kalita [2011]	Clustering	Dominio de datos abierto.
Long et al. [2011]	Clustering	Dominio de datos abierto; Optimiza distintos criterios para seleccionar mensajes
Sankaranarayanan et al. [2009]	Clustering	Resumenes en tiempo real; dominio de datos abierto; eventos noticiosos.
Metzler et al. [2012]	Probabilísticos	dominio de datos abierto; optimiza distintos criterios para seleccionar mensajes.
Bian et al. [2015]	Probabilísticos	considera contenido multimedia; dominio de datos abierto; considera distintos aspectos para evaluar.

Tabla 3.1: Tabla resumen trabajos relevantes

A continuación se presentan los trabajos asociados a está área, agrupados según las tres categorías presentadas. Además, por cada uno de las categorías se explica en detalle un método representativo.

3.1.1. Enfoques basados en grafos

Los trabajos que utilizan estos métodos Schinas et al. [2016], Sharifi et al. [2010], Xu et al. [2013] buscan aprovechar las distintas relaciones entre los mensajes y usuarios de la red social para modelar una estructura de grafo, por ejemplo, construyendo un grafo a partir de las interacciones entre usuarios o utilizando las relaciones entre los mensajes también.

Uno de los primeros trabajos relacionados con la generación de resúmenes automáticos a partir de tweets es el realizado por Sharifi et al. [2010] denominado Phrase Reinforcement Algorithm. Este algoritmo se basa en dos aspectos importantes de los mensajes de redes sociales, la primera es que los usuarios usan palabras similares para describir un tema en particular. El segundo aspecto corresponde a que en muchos casos los usuarios republican los mensajes más populares de un evento, esto aumenta el calce entre las palabras de los mensajes. El objetivo del algoritmo es producir la frase que tiene mayor calce con el conjunto

original de mensajes. Esto porque la cantidad de veces que se repite una frase es un buen indicador de su importancia. Para lograr esto el algoritmo toma como entrada 2 elementos; el primero una frase inicial o raiz, que puede corresponder a un trending topic, o una palabra clave seleccionada por el usuario; el segundo es un conjunto de tweets que contengan la frase inicial. Este método fue escogido para la generación de resúmenes en los experimentos, a continuación, se explica cada una de cuatro etapas que lo conforman:

1. **Construcción grafo de palabras** Primero se busca armar un grafo el cual muestra la ocurrencia de palabras y frases que están antes y después del nodo raíz (palabra clave). Para armar este grafo se divide en 2 sub-grafos, el derecho, el cual contendrá las frases que aparecen a la derecha del nodo raíz y el sub-grafo izquierdo el cual a su vez contendrá las frases a la izquierda del nodo raíz. Para armar el sub-grafo derecho, primero se consideran las palabras inmediatamente a la derecha del nodo raíz, estas palabras son agregadas como nodos y además se incluye un contador de cuantas veces apareció dicha palabra en cada mensaje, cada palabra es pasada a minúscula y los carácteres no alfanuméricos son eliminados. Son excluidas las palabras cuyo contador sean menor a 2. Luego se repite el proceso para cada uno de los nodos, ahora considerando como nodo raíz, los nodos agregados en el paso anterior. El proceso para general el sub-grafo izquierdo es análogo. Este proceso continua hasta que se han procesado todos los mensajes.

2. **Pesando nodos individuales** Una vez generado ambos subgrafos ahora se le asigna un peso a cada nodo, para lo cual se utiliza la siguiente formula.

$$\text{Weight(Nodo)} = \text{Count(Nodo)-Distance(Nodo)} \cdot log_b Count(Nodo) \tag{3.1}$$

   En este caso *Distance*, es la distancia desde el nodo actual al nodo raíz. *Count* es la cantidad de veces que se repite dicha palabra. El valor de b se puede modificar para establecer una preferencia entre resúmenes más cortos o más largos.

3. **Generación del resumen** Ahora para generar un resumen, se busca la frase con mayor traslape en el grafo. Para realizar esto se generan dos resúmenes parciales.

   El primer resumen se genera a partir del nodo raíz y buscando el camino con mayor peso que termine en un nodo que se encuentre al costado izquierdo del nodo raíz. Este camino constituye el resumen parcial izquierdo, ahora para generar el resumen parcial derecho, se considera todo el resumen parcial izquierdo como un solo nodo, el cual pasará a ser considerado el nodo raíz, a continuación, se busca el camino con mayor peso que se encuentre desde el nodo raíz, este camino será el resumen parcial derecho.

   Finalmente, el resumen generado por el algoritmo es la concatenación del resumen parcial derecho con el resumen parcial izquierdo.

4. **Post-procesamiento** El resumen obtenido no necesariamente tiene la coherencia y cohesión necesarias para ser leído, debido al preprocesamiento realizado. Para solucionar esto, se busca dentro de los mensajes originales el que mejor calce con el resumen generado, para obtener dicho mensaje se divide el resumen generado en tokens, luego se busca en orden la ocurrencia de cada token en una frase, permitiendo además cierto número de caracteres no alfanuméricos entre cada token, una vez que se obtiene un

mensaje que calce con los tokens, este pasa a ser el resumen final.

A modo de ejemplo si tomamos el término “ted kennedy” y las siguientes frases como input: “ tragedy: Ted Kennedy died today of cancer”; “Ted Kennedy died today”; “Ted Kennedy was a leader”; “Ted Kennedy died at age 77”, producirán el grafo que se muestra en la figura 3.1. Si se busca la frase con mayor calce dentro del grafo, corresponde a “Ted Kennedy died today”, y para este caso el resumen estara compuesto por la frase única con más peso que corresponde a “A tragedy: "Ted Kennedy died today of cancer”.

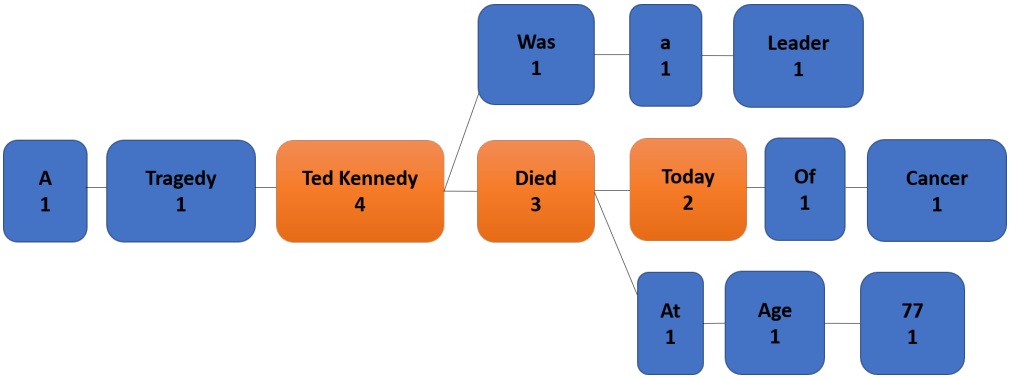

Figura 3.1: Ejemplo del grafo de palabras obtenido para un conjunto de mensajes, asociados al término “ted kennedy”

El trabajo de Sharifi et al. [2010], fue considerado importante, ya que es utilizado por otros métodos como baseline para comparación; Trabaja sobre un conjunto de eventos de distintas características y es simple de implementar. Un trabajo importante es el propuesto por Schinas et al. [2016], llamado *MGraph*, este método busca hacer una selección de imágenes y mensajes compartidos en Twitter para resumir un evento. Para obtener esta selección, se construye un multigrafo que considera; características de similitud del texto, similitud de las imágenes y relaciones sociales entre mensajes, para finalmente asignarle un puntaje a cada mensaje. Este trabajo fue utilizado como método para generar los resúmenes en este trabajo. A continuación, se explica en cada una de sus etapas.

Representación de datos Según los autores los mensajes publicados en una red social pueden ser vistos como elementos multimodales, descritos por la siguiente tupla (id, C, u, t_s, SI), en donde id, representa el identificador único del mensaje; C, corresponde al contenido del mensaje, el cual puede contener 2 partes C_{vis}, correspondiente al contenido visual y C_{text} correspondiente al contenido textual del mensaje; u, es el usuario que ha publicado el mensaje; t_s corresponde a la fecha de publicación del mensaje y SI corresponde a las interacciones sociales que ha tenido dicho mensaje. El contenido visual del mensaje está representado con descriptores SURF. Por otra parte, el contenido textual es representado por vectores tf-idf, pero con 2 modificaciones, primero el valor de idf es calculado sobre el conjunto total de mensajes y no sobre cada mensaje individual y segundo se identifican entidades nombradas dentro de los mensajes y su valor es amplificado por una constante b. Para el caso de las interacciones sociales se consideran las respuestas que ha tenido el mensaje, los repost y las menciones de usuario, se almacena el id del post original en caso de que sea una respuesta.

Filtrado En esta etapa se busca limpiar y descartar todos los posibles mensajes que no aporten información relevante. Para esto primero se filtra por su contenido textual, se descartan los mensajes cuyo texto contenga menos de 6 términos, contengan más de 3 hashtags, más de 3 menciones a usuarios o más de 3 urls. Además, se eliminan mensajes que estén mal redactados, considerando elementos que tengan al menos una oración compuesta por sustantivo seguido de un verbo. El segundo tipo de filtro es un filtro aplicado sobre las imágenes. Primero se eliminan aquellas cuyas dimensiones sean menos de 200px. Luego los autores entrenan un clasificador, utilizando SVM, para dividir las imágenes en 4 categorías *real photo*, *meme*, *screenshoot*, *heavy text* para finalmente solo mantener aquellas etiquetadas como *real photo*.

Generación del multígrafo Los mensajes restantes son usados para construir un multígrafo, $G_M = (V, E_{txt}, E_{\text{vis}}, E_{soc}, E_{time})$ donde cada vértice representa un mensaje. E_{txt} es un conjunto de arcos expresando la similitud coseno entre el contenido textual de los mensajes, un arco es agregado solo si la similitud es mayor a un cierto umbral; E_{vis} es un conjunto de arcos que representa la similitud entre las imágenes de un mensaje. La similitud en este caso es calculada con la distancia L_2 sobre los vectores SURF (Bay et al. [2006]) de cada imagen, al igual que en la similitud textual, el arco solo es agregado si la similitud es mayor que un umbral especifico. E_{soc} son los arcos que representan las interacciones sociales entre mensajes. Se agrega un arco entre 2 nodos si uno es respuesta del otro o si uno es un repost del otro. E_{time} es el conjunto de arcos que representa la proximidad temporal entre los mensajes, la proximidad temporal entre los vértices es calculada con la siguiente ecuación.

$$TS(t_j, t_k) = \text{e}xp(\frac{-|(t_j - t_k)|^2}{2\sigma^2}) \tag{3.2}$$

El valor de σ debe ser ajustado para cada evento, en el caso de eventos más largos, se requiere un valor de σ mayor.

De-duplicación de contenido Para mejorar la diversidad y relevancia de los elementos dentro del resumen es importante eliminar la mayor cantidad posible de duplicados. En el caso del contenido textual se descartan todos los mensajes que sean un repost de otro mensaje. Para el caso del contenido visual de un mensaje un usuario puede postear la misma imagen o una muy similar, para manejar esto los autores utilizan un método llamado *Clique Percolation Method* o CPM. Primero considerando el subgrafo que contiene solo el contenido visual del grafo se filtran todos los arcos que estén bajo un umbral θ y luego utilizan CPM para descubrir cliques dentro de este grafo. Los cliques resultantes son representados como un solo mensaje, en específico el clique mc es representado por la tupla (M_{mc}, C, t_s, p), donde M_{mc} es el conjunto de post que forman el clique, C es una representación del contenido agregado de los mensajes, t_s es el tiempo de publicación promedio de los mensajes y p es la suma de reposts de los mensajes. El contenido textual es representando por la suma de los vectores tf-idf de los mensajes, ahora para el caso del contenido visual, este es representado por un descriptor que se obtiene al agregar todos los descriptores de las imágenes. Luego de detectar los cliques, los mensajes en dichos cliques son reemplazados por la representación antes explicada y los nuevos arcos del grafo son actualizados.

Detección de tópicos Para detectar los distintos tópicos de un evento, los autores utilizan

Structural Clustering Algorithm for Networks (SCAN). Este algoritmo es aplicado sobre un grafo $G = (V, E)$, donde los vértices en V corresponden al conjunto de post y cliques filtrados en los pasos anteriores. Los arcos representan cuando dos mensajes hablan sobre el mismo tópico. En particular para agregar un arco en el grafo, primero se considera si existen arcos temporales y de contenido, es decir dos nodos se conectan si existen una proximidad temporal entre ellos y existe similitud de contenido. Además, se agregan arcos si uno de los mensajes es respuesta del otro, independiente de las otras características. Una vez construido este grafo, se aplica el algoritmo SCAN, para identificar sub-grafos densos de mensajes. Estos subgrafos representan los tópicos en la colección de mensajes, una vez detectados los T tópicos, se calcula un vector v_{i}^{tp} por cada tópico $tp_{\mathrm{i}} \in T$ este vector es una representación del contenido textual del tópico y es calculado de la misma forma descrita en la sección de detección de cliques. En el caso del contenido visual no se calcula un centroide que lo represente, debido a que las imágenes de un mismo tópico pueden ser muy distintas. En el caso que algunos mensajes no quedaran asignados a ningún tópico estos mensajes no son descartados, ya que pueden seguir siendo relevantes para el evento, por ejemplo, los mensajes con imágenes pueden tener poco contenido textual y las imágenes ser muy distintas entre sí para un mismo tópico, pero seguir siendo relevantes, es por esto que se mantienen y forman clúster con un solo ítem. Estos mensajes pueden ser divididos en dos categorías *outliers* y *hubs*. Los *outliers* corresponden a mensajes que no están conectados a ningún otro clúster y los *hubs* corresponde a mensajes que están conectados a más de un clúster, en ambos casos estos mensajes son conservados y utilizados más adelante en el proceso de ranking.

Selección de mensajes y ranking La meta principal es calcular un puntaje general para cada mensaje y clique, con el fin de mostrar los más relevantes. Dos factores son considerados al momento de calcular el puntaje, la atención que atrajo en la red social y la importancia del tópico asociado a ese mensaje. El puntaje de un mensaje o clique está representado por:

$$S(m) = S_{soc}(m) \cdot S_{cov}(m) \tag{3.3}$$

Donde S_{soc} es el puntaje de popularidad del mensaje. Es calculado se la siguiente forma

$$S_{soc}(m) = \log(p + \lambda) \tag{3.4}$$

Donde p representa el número de repost de un mensaje y λ es un parámetro para garantizar valores distintos de cero.

Por otra parte S_{cov} es el puntaje asociado al tópico al cual pertenece el mensaje y es calculado de la siguiente forma.

$$S_{cov} = \cos(u_m, v_{\mathrm{i}}^{tp}) S(tp_{\mathrm{i}}) \tag{3.5}$$

Donde $S(tp_{\mathrm{i}})$ esta definido por:

$$S(tp_{\mathrm{i}}) = D_{\mathrm{i}} \cdot \exp\left(\frac{|M_{\mathrm{i}}|}{\max_{k \in T} |M_k|}\right) \tag{3.6}$$

y el termino D_{i} es calculado como

$$D_{\mathrm{i}} = \frac{2|E_{\mathrm{i}}|}{|V_{\mathrm{i}}||V_{\mathrm{i}} - 1} \tag{3.7}$$

El primer término de la ecuación 3.5 representa la similitud textual entre el mensaje y el centroide del tópico. El segundo término representa la importancia del tópico S, los tópicos más grandes y densos reciben un puntaje mayor, para determinar esto se considera el subgrafo producido en la etapa anterior por SCAN y la densidad del tópico tp_{i} está dada por el termino D_{i}, en donde $|E_{\mathrm{i}}|$ es el número de arcos y $|V_{\mathrm{i}}|$ es el número de vértices. $|M_{\mathrm{i}}|$ representa la cantidad de mensajes asociados al tópico i.

Ranking de imágenes y diversificación Existen imágenes que atraen mucha atención social durante un evento, pero no aportan información relevante sobre el evento en sí. Para solucionar esto, se propone el concepto de *especifidad de la imagen (S_{spec})*. Para poder determinar este valor para una imagen I, se utiliza un esquema similar a tf-idf.

$$S_{spec}(I) = \log \frac{|T|}{|T_I|} \tag{3.8}$$

En donde $|T|$ es el número de tópicos del evento y $|T_I|$ es el número de tópicos en donde aparece la imagen I. Para poder estimar $|T_I|$ de una imagen y su cuasi-duplicados, se utiliza nuevamente la técnica CPM descrita anteriormente. Luego para las imágenes que no forman cliques, se revisa si están contenidas en alguno de los *hubs* encontrados por el algoritmo SCAN, en la fase de detección de tópicos, revisando con cuantas comunidades está conectado, pero solo considerándolas si son superiores a un umbral particular.

Finalmente, el puntaje de una imagen corresponde al producto entre S_{spec} y S explicado en la ecuación 2.2. Para evitar la redundancia al momento de seleccionar las imágenes, se aplica DivRank, sobre el conjunto de imágenes. Esta lista ordena de imágenes con sus respectivos mensajes corresponde al resumen final mostrado al usuario.

En la figura 3.2 se muestra un diagrama resumen con cada una de las etapas del método *MGraph*.

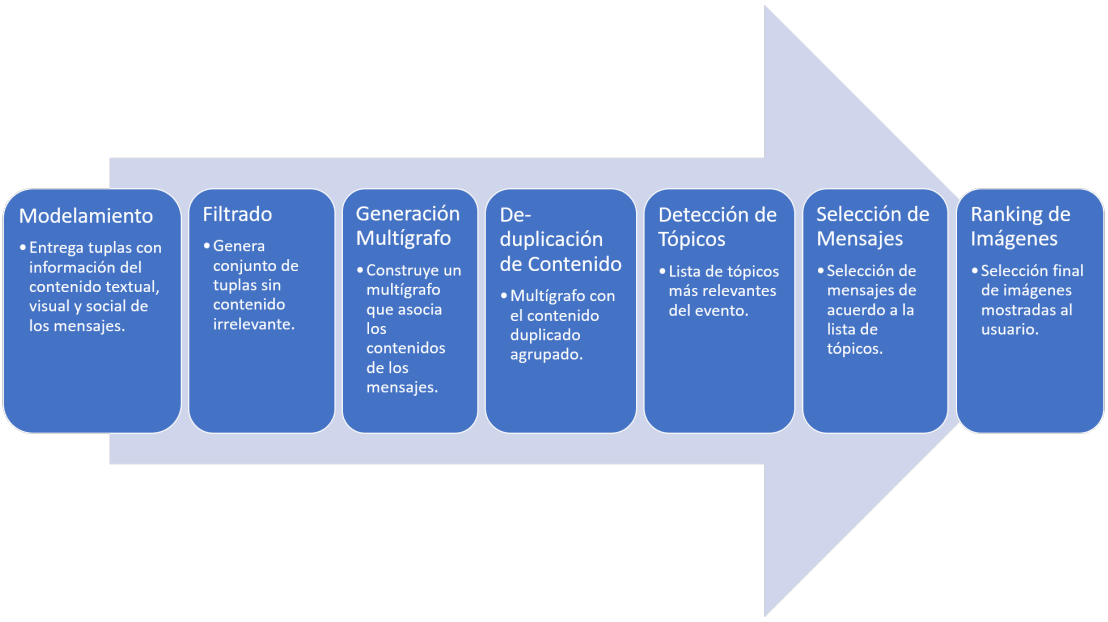

Figura 3.2: Resumen de etapas para MGraph

En Xu et al. [2016], se propone utilizar una red de redes, esto para aprovechar toda la información disponible en los tweets (texto, imágenes y links de noticias), donde cada capa de la red representa un tipo de información distinta. Generando una red solo para el contenido textual de los mensajes; otra red para las imágenes presente en los mensajes y otra red para las noticias compartidas por los mensajes. Los arcos dentro de una red son calculados en base a medidas de similaridad entre los elementos. También se agregan arcos entre las redes a partir de las urls presentes en los mensajes. Una vez construida esta representación busca seleccionar k tuplas que maximicen la diversidad de los contenidos, maximicen la importancia del contenido extraído y que estén ordenados de forma temporal.

La gran mayoría de los métodos corresponden a esta categoría, esto se debe principalmente a que las interacciones en las redes sociales son fácilmente modeladas utilizando un grafo. Además, al utilizar un grafo se pueden incluir distintos tipos de relaciones, como por ejemplo, relaciones temporales, relaciones de similitud semántica, etc. La principal dificultad que presentan los trabajos en esta categoría es: Si se ejecutan sobre un gran número de mensajes, su tiempo de ejecución aumenta, ya que debe calcular el grafo con sus respectivos arcos para cada uno de los elementos considerados, es decir, el tiempo de ejecución aumenta proporcionalmente a la cantidad de elementos por la cantidad de arcos del grafo. Esto impide que este tipo de método puede ser utilizado en tiempo real.

3.1.2. Métodos enfocados en clustering

Existen distintos trabajos que utilizan clustering como su técnica principal Sankaranarayanan et al. [2009], Inouye and Kalita [2011], Steinbach et al. [2000], Long et al. [2011]. La idea principal consiste en utilizar distintas técnicas de clustering para detectar los subeventos existentes y luego seleccionar elementos representantes de cada clústers. Se pueden utilizar distintas técnicas de clustering, dependiendo del tipo de resumen que se quiera generar. Por ejemplo, si se quiere generar un resumen en tiempo real, a medida que van apareciendo nuevos mensajes, es preferible utilizar clustering incremental, ya que no se sabe a priori el número de clústers existentes. Este enfoque es abordado por Sankaranarayanan et al. [2009], en ese trabajo buscan extraer tweets relevantes para noticias de último minuto. Para realizar esto los autores proponen utilizar clustering incremental para segmentar por tópicos los tweets en la medida que estos van siendo publicados. Posteriormente, utilizando características geográficas de los tweets, se ordenan las ubicaciones más relevantes y se seleccionan los mensajes en base a su contenido.

El trabajo propuesto por Inouye and Kalita [2011], utiliza mensajes asociados a 50 trending topics, recolectados por 5 días. Lo que busca este trabajo es dado un evento extraer los distintos subeventos, usando clustering, para posteriormente por cada clúster seleccionar el mensaje que obtenga el mayor puntaje de acuerdo a una métrica definida por los autores llamada *Hybrid tf-idf*. Este algoritmo cuenta con 3 pasos:

Filtrado Se eliminan de los mensajes todas las url, menciones de usuario, tags con el fin de mantener solo palabras en inglés.

Clustering Aplicar bisecting k-means sobre el conjunto de mensajes para agrupar todos los mensajes que tienen contenido similar.

Hybrid tf-idf Por cada clúster generado en la etapa anterior y para cada mensaje se calcula el valor de la métrica *Hybrid tf-idf*, basada en el tf-idf original, pero buscando compensar el largo limitado de los mensajes, la cual está definida para un mensaje s como:

$$W(s) = \frac{\sum_{\mathrm{i}=0}^{n} tf(W_{\mathrm{i}}) \cdot \mathrm{id}f(W_{\mathrm{i}})}{nf} \tag{3.9}$$

Donde **nf** es un término de normalización, n es el número de palabras de un mensaje;W_{i} corresponde al i-ésimo término del mensaje s; tf está definido como el número de veces que aparece el término en todos los mensajes, dividido por la cantidad total de palabras y *idf* está definido como número de mensajes dividido en mensajes en los cuales aparece el término. Si se usuara tf-idf normal, se tendria que cada mensaje sería un documento y al ser tan cortos el valor de tf sería muy pequeño con respecto a idf, en cambio, si ahora se considera todos los mensajes de un evento como un documento, entonces el valor de idf sería muy pequeño, ya que solo habría un documento.

En Long et al. [2011], detectan los eventos utilizando los hashtags mencionados en las redes sociales, comparando la frecuencia con la cual son compartidas. Para luego generar un grafo con estas palabras donde los vértices son las palabras detectas y se agrega un arco, cada vez que la palabra w_{i} y la palabra w_j, aparecen juntas en un mismo mensaje. Posteriormente, se aplica clustering jerárquico sobre este grafo, hasta obtener k clúster. Luego para obtener el resumen se busca generar un conjunto que maximice la cobertura, considerando la similitud entre los elementos y sus tiempos de publicación.

En el caso de los métodos basados en clustering, estos son más simples de implementar, ya que una vez extraídos los eventos, se aplica algún algoritmo de clustering y se seleccionan los mensajes. La eficiencia de los algoritmos de clustering además permite que estos métodos puedan ser utilizados para generar los resúmenes en tiempo real a medida que surgen los mensajes, este es el caso del trabajo de Sankaranarayanan et al. [2009].

3.1.3. Enfoques basados en modelos probabilísticos

En este caso, los trabajos asociados Chakrabarti and Punera [2011], Bian et al. [2015], Metzler et al. [2012], Meng et al. [2012], utilizan distintos modelos probabilísticos para detectar los subtópicos del evento. Para posteriormente seleccionar un conjunto de mensajes según una combinación de criterios especificados por cada autor. Se pueden utilizar distintos modelos para generar los resúmenes

Uno de los primeros trabajos en utilizar este enfoque es el propuesto por Chakrabarti and Punera [2011]. Los autores buscan resumir eventos deportivos utilizando Twitter. En este caso los autores buscan aprovechar la estructura definida y recurrente que poseen los eventos deportivos, por ejemplo, un partido siempre tiene los mismos tipos de acontecimientos (goles, faltas, penal, entretiempo). Ellos proponen utilizar Hidden Markov Models para detectar y modelar los peaks de actividad durante el evento, estos peaks corresponden a los hechos más relevantes del juego, como una anotación o expulsión. Una vez detectados los subeventos, se escogen k representantes, estos representantes son escogidos como los mensajes que tiene la menor distancia de similitud con respecto a todos los demás mensajes del subevento.

En el modelo propuesto por Bian et al. [2015], buscan producir resúmenes tanto de imágenes como de texto en redes sociales. Proponen un modelo basado en LDA Blei et al. [2003], llamado CMLDA, con este modelo ellos buscan caracterizar los subtopicos existentes en un evento en particular. El modelo propuesto tiene como diferencia con respecto al LDA estándar que los mensajes solo pueden pertenecer a un solo subtópico, además de considerar tanto características textuales como visuales de los mensajes. Este método primero particionara los mensajes en K grupos donde cada grupo tendrá una parte visual (imágenes que fueron asociadas a ese grupo) y una parte textual (texto que fue asociado a ese grupo), cada uno de estos grupos representa un subevento. Luego para generar el resumen final, se genera un resumen del contenido textual y otro resumen con las imágenes, para la generación del resumen los autores consideran *cobertura, diversidad* y *relevancia*. Este método utiliza eventos noticiosos escogidos al azar a partir de un conjunto de trending topics recolectados durante un mes utilizando la api de twitter.

Otro trabajo es el desarrollado por Metzler et al. [2012], en este caso dado un término en particular q, por ejemplo, "terremoto".Se busca generar una *representación estructurada del evento*, la cual consiste en hora de inicio, duración y un conjunto de mensajes que resumen el evento, en la figura 3.3 se muestra un ejemplo para la palabra earthquake. Para generar está representación se realizan los siguientes pasos:

Expansión de query Dado que pueden existir mensajes asociados al término buscado, que no contengan el término, por ejemplo mensajes que tengan en vez de la palabra "terremoto", contengan la palabra "temblor". Por lo cual se deben expandir los terminos de búsqueda. Para esto se realiza lo siguiente. Primero se seleccionan N intervalos de tiempo donde el término q haya sido discutido, estos intervalos son ordenados de acuerdo a la proporción de mensajes que contienen a q. En cada uno de los intervalos se calcula un puntaje de $burstiness$ para cada termino de cada mensaje del intervalo. Este puntaje busca cuantificar que tan popular fue un término en un intervalo, está representado por la probabilidad que un termino ocurra en este intervalo de tiempo, versus la probabilidad que el mimsmo término ocurra en cualquier intervalo . El puntaje de $burstiness$ está definido por la ecuación 3.10

$$burstiness(w, TS_{\mathrm{i}}) = \frac{P(w|TS_{\mathrm{i}})}{P(w)} \tag{3.10}$$

Donde $P(w|TS_{\mathrm{i}})$ esta definido por la siguiente ecuación.

$$P(w|TS_{\mathrm{i}}) = \frac{tf_{w,TS_{\mathrm{i}}} + \nu\frac{tf_w}{N}}{|TS_{\mathrm{i}}| + \nu} \tag{3.11}$$

El término $P(w) = \frac{tf_w + K}{N + K|V|}$. En este caso $tf_{w,TS_{\mathrm{i}}}$ es la frecuencia del término w en el intervalo TS_{i}; tf_w es la frecuencia de w en todo conjunto de mensajes; $|TS_{\mathrm{i}}|$ es la cantidad de terminos en el intervalo; V es el tamaño del vocabulario; K y ν son términos definidos por los autores. Una vez calculados los puntajes de cada termino para cada intervalo, el puntaje final de un término (β_w) es el promedio de sus puntajes en cada intervalo. Luego los k términos con el mayor puntaje son seleccionados.

Ranking temporal Una vez obtenido el conjunto expandido q' con sus respectivos puntajes. Se seleccionan los D intervalos que tengan el puntaje más alto, posteriormente los

intervalos continuos hasta obtener un solo intervalo, que representa toda la discusión del evento . El puntaje de un intervalo con respecto q', busca asignar un puntaje más alto a aquellos intervalos que cubren más términos del conjunto q'. Este puntaje esta definido por:

$$s(q, TS) = \sum_{w \in q} \beta_w tf_{w,TS} \tag{3.12}$$

Resumen temporal El último paso es seleccionar mensajes que representen al intervalo de tiempo de la etapa anterior. Para esto se seleccionan los k mensajes que tengan mayor puntaje de relevancia para el conjunto q'. Este puntaje para un mensaje M y el conjunto q' esta definido por

$$s(q, M) = \sum_{w \in q} \beta_w \log P(w|M) \tag{3.13}$$

Donde P(w|M) es la función de relevancia, propuesta por Ponte and Croft [1998]

July 16 2010 at 17 UTC, for 11 hours
Summary tweets:
Ok a 3.6 “rocks” nothing. But boarding a plane there now, Woodward ho! RT @todayshow: 3.6 magnitude #earthquake rocks Washington DC area.
RT @fredthompson: 3.6-magnitude earthquake hit DC. President Obama said it was due to 8 years of Bush failing to regulate plate tectonic ...
3.6-magnitude earthquake wakes Md. residents: Temblor centered in Gaithersburg felt by as many as 3 million people... http://bit.ly/9iMLEk

Figura 3.3: Ejemplo de representación estructurada para la palabra earthquake

Este método también utiliza eventos extraídos al azar utilizando la API de Twitter, extrayendo el 1 % de los mensajes publicados por día durante un periodo de 6 meses.

Para los análisis y comparaciones posteriores, no se considero ninguno de los métodos que utilizan modelos probabilísticos. Estos trabajos no fueron considerados, porque ninguno cumplía con los requisitos. Por ejemplo, el trabajo de Metzler et al. [2012], no considera distintos criterios ni para seleccionar mensajes ni para evaluar los resultados. En el caso del trabajo realizado por Bian et al. [2015] al considerar el contenido multimedia de los mensajes el desarrollo de dicho método se complejizo y no pudo ser implementado.

3.1.4. Otros enfoques

Aunque la gran mayoría de los trabajos realizados en esta área pertenece a una o más de las categorías antes descritas, existen otros trabajos que han buscado un enfoque distinto

para abordar esta problemática, en particular existen trabajos que buscan modelar este problema, como un problema de optimización en donde se define una función para cuantificar un determinado conjunto criterios, por ejemplo *diversidad* y *relevancia*, para posteriormente escoger los mensajes que maximicen estas funciones.

Por ejemplo, en Štajner et al. [2013], abordan el problema de generar resúmenes de noticias comentadas en la red social, para esto buscan optimizar maximizar una función, que acorde a los autores mide la utilidad del conjunto de mensajes, es decir, buscan el conjunto de mensajes que provea la máxima utilidad, en el caso de los autores definen la utilidad de un mensaje como una ponderación de características sociales, características de su contenido y características del autor del mensaje.

En el trabajo realizado por Yan et al. [2011], los autores buscan generar líneas de tiempo que resuman un evento. Para esto abordan un enfoque similar al anterior con la diferencia que ahora su función de utilidad está definida como una combinación lineal de funciones definidas por los autores, donde cada una de estas funciones busca representar un criterio especifico, como cobertura, relevancia, coherencia y diversidad.

Capítulo 4

Evaluación de resúmenes

Una vez generado un resumen de forma automática, es importante comprobar que este resumen es una correcta representación de la fuente original.

La evaluación de un resumen es una tarea muy compleja, ya que no existe un criterio único totalmente confiable, para estimar la calidad de un resumen en todos los casos. Esto debido principalmente a que si dos personas generan un resumen para el mismo documento, estos resúmenes serán distintos, pero ambos pueden ser buenas representaciones del conjunto original. Esto puede deberse a diversos motivos, por ejemplo, considerar distintos niveles de abstracción; darle más importancia a un tema por sobre otro; el tipo de usuario al cual está enfocado el resumen; el momento en el cual se genero el resumen. En el caso de resúmenes generados de forma automática esto también se aplica ya que distintos sistemas, producirán distintos resúmenes, de acuerdo a sus configuraciones y parámetros.

En este capítulo se presentan distintas formas para evaluar un sistema de generación de resúmenes mostrando, detallando las categorías existentes y las métricas asociadas.

4.1. Técnicas de evaluación

En esta área existen dos enfoques principales para realizar la evaluación, propuestos por Jones and Galliers [1995]. La primera es la evaluación de tipo *intrínseca*; en este tipo de evaluaciones lo que se busca medir es la calidad del resumen como tal, solo considerando sus propiedades internas, considerando distintos criterios específicos. El segundo enfoque, para realizar la evaluación de un resumen automático, es la evaluación *extrínsica*; está metodología tiene como objetivo evaluar cómo afecta el resumen a la realización de una tarea específica, como por ejemplo responder un cuestionario para medir el nivel de conocimiento que logran adquirir los usuarios después de leer el resumen.

En la figura 4.1, se muestra un resumen de la categorización utilizada para presentar las métricas existentes en la evaluación de resúmenes automáticos.

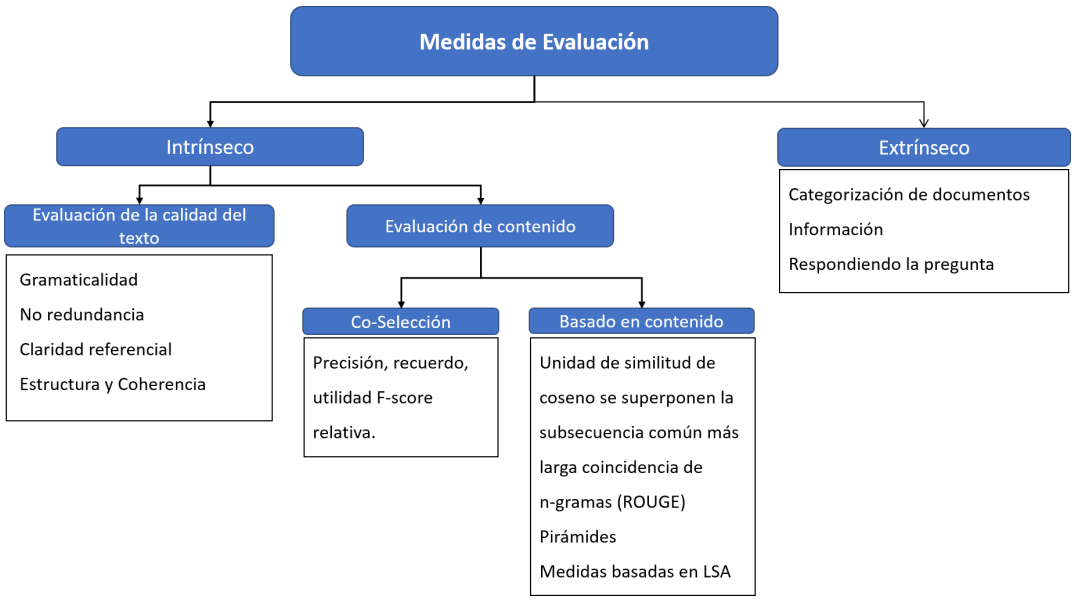

Figura 4.1: Categorización de métricas para resúmenes automáticos

4.1.1. Evaluación intrínseca

La evaluación intrínseca, se centra en evaluar la calidad del contenido textual presente en el resumen. De acuerdo a Steinberger and Ježek [2012], puede ser dividida en tres enfoques. El primero, evaluación de calidad de texto; el segundo medidas de co-selección y el tercero medidas basadas en contenido. A continuación se explica cada enfoque y se mencionan los trabajos en los que fueron utilizados:

Evaluación de calidad de texto

En este primer tipo de evaluación, los criterios buscan que el resumen generado sea comprensible para una persona. Schinas et al. [2016], McParlane et al. [2014], Xu and Lu [2015], sin importar si su contenido es relevante o no, existen diversos criterios de evaluación:

- Gramática, el resumen no debería contener errores de puntuación o palabras mal escritas.
- No-Redundancia, el resumen no debe contener información redundante.
- Claridad de Referencia, los sustantivos, comunes o propios deben estar claramente referenciados en el resumen.
- Coherencia y estructura, la estructura del resumen debe ser correcta y las frases deben ser coherentes.

El principal problema con estos criterios es que no son simples de medir de forma automática. Lo cual obliga a realizar una evaluación manual con usuarios, en la cual el grupo de usuarios asigna una nota según cada criterio.

Este tipo de evaluación es aplicado por Schinas et al. [2016] y por McParlane et al. [2014], en la cual se les pide a un conjunto de usuarios que etiqueten en una escala de 1 a 5 respecto a que tan importante es un mensaje de Twitter dado un evento en particular. Luego a partir de esa información son capaces de medir cuantos de los mensajes marcados como relevante o importantes fueron seleccionados por sus respectivos métodos.

Otro enfoque de esta metodología es preguntar directamente a los usuarios sobre la calidad del resumen. Preguntando, por ejemplo, si el resumen generado cubre distintos aspectos del evento o si el resumen ayuda a la comprensión del evento, este enfoque es utilizado por los autores de Xu and Lu [2015].

Medidas de Co-selección

Se busca evaluar que el contenido del resumen sea importante, o que esté relacionado con los documentos originales que fueron resumidos. La forma más común de realizar esta evaluación es tener dos conjuntos de resúmenes. El primero, son los generados de forma automática por el sistema, y el segundo son resúmenes generados por usuarios de forma manual. Una vez que se tienen los dos conjuntos de resúmenes, se busca comparar la cantidad de frases en común. Este tipo de evaluación solo se puede usar en sistemas que sean extractivos, ya que mide el calce de frases completas entre ambos conjuntos.

Precisión Basado en las métrica original de recuperación de la información. Corresponde a el número de frases que están presentes en ambos conjuntos, dividido por el número de frases presentes en el conjunto generado automáticamente. Esta definido como:

$$P = \frac{|Frases\ en\ el\ resumen\ \text{ideal}| \cap |Frases\ en\ el\ resumen\ automtico|}{|Frases\ en\ el\ resumen\ automtico|} \tag{4.1}$$

Recall Basado en las métrica original de recuperación de la información. Corresponde a el número de frases que están presentes en ambos conjuntos dividido por el número de frases presenten en el conjunto de resúmenes ideales. Esta definido como:

$$R = \frac{|Frases\ en\ el\ resumen\ \text{ideal}| \cap |Frases\ en\ el\ resumen\ automtico|}{|Frases\ en\ el\ resumen\ \text{ideal}|} \tag{4.2}$$

F-Score Es una medida compuesta que considera tanto presicion como recall. Esta definido como:

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{4.3}$$

Utilidad Relativa Propuesta por Radev et al. [2004], consiste en que cada juez le otorga un puntaje a cada frase de los documentos originales. Este puntaje indica el grado al cual debería pertenecer dicha frase al resumen, según el juez, se denomina utilidad de la frase u. A partir de esto se puede obtener la siguiente métrica.

$$RU = \frac{\sum_{j=1}^{n} \delta_j \sum_{\mathrm{i}=1}^{N} u_{\mathrm{i}j}}{\sum_{j=1}^{n} \varepsilon_j \sum_{\mathrm{i}=1}^{N} u_{\mathrm{i}j}} \tag{4.4}$$

Donde u_{ij}, es la utilidad de la frase j dada por el juez i; ε_j es 1 para las primeras e frases, de acuerdo a la suma de sus utilidades asignadas por todos los jueces, para las demás es cero y δ_j es 1 para las primeras e frases extraídas por el sistema, para las demás es cero.

Estas medidas solo son útiles si la metodología para generar el resumen es extractivo, es decir, sacar frases textuales del conjunto original, ya que buscaran el calce exacto entre los términos.

Medidas basadas en el contenido

El principal problema con las métricas de co-selección, es que solo cuentan como un acierto si toda la frase coincide, ignorando el hecho que dos frases pueden contener la misma información escrita de forma distinta. Por esto las métricas mostradas anteriormente no pueden utilizarse para sistemas de resúmenes abstractivos. En cambio las medidas basadas en contenido comparan las palabras de forma individual en vez de la frase completa.

Similitud Coseno De acuerdo a Salton [1989] es una medida de similitud entre dos vectores distintos de cero. La similitud coseno entre la representación vectorial del resumen X y del resumen ideal Y está definida de la siguiente forma:

$$\cos(X, Y) = \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\|\|\vec{Y}\|} \tag{4.5}$$

En el caso de resúmenes automáticos, si el resumen tiene una alta similaridad con los resúmenes humanos, se puede deducir que su contenido es similar.

Indice de Jaccard Esta métrica es utilizada para medir la similaridad y diversidad de dos conjuntos finitos de elementos. Está definida como el tamaño de la intersección dividido por el tamaño de la unión de dichos conjuntos.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.6}$$

En este caso particular esta métrica es bastante útil, ya que si se tienen por ejemplo dos conjuntos de palabras con el índice de Jaccard podemos saber cuántos términos en común tienen y determinar si cubren temas similares.

Subsequencia más larga Otra métrica basada en contenido es "Subsequencia más larga", propuesta por Radev et al. [2003]. Está definida como:

$$LCS(X, Y) = \frac{length(X) + length(Y) - \text{edi}t(X, Y)}{2} \tag{4.7}$$

$LCS(X, Y)$ es el largo de la subsequencia común más larga entre X e Y, $length(X)$, representa el largo del texto X y edi$t(X, Y)$.

ROUGE Conjunto de métricas especificas para evaluar resúmenes, propuestas por Lin [2004], para facilitar la evaluación de resúmenes generados automáticamente. Consiste en buscar el calce de n-gramas dentro de un conjunto de texto, donde uno de los textos

es un resumen generado automáticamente y los demás son resúmenes de referencia que actúan como gold standard

$$\text{ROUGE-N} = \frac{\sum_{S \in R} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in R} \sum_{gram_n \in S} Count(gram_n)} \tag{4.8}$$

Donde R es el conjunto de los resúmenes de referencia, n es el largo de los n-gramas buscados y $Count_{match}()$ es el número máximo de n-gramas que co-ocurren tanto en el resumen candidato como en los resúmenes de referencia.

En el caso de resúmenes que utilizan redes sociales es mejor usar ROUGE-1, es decir buscar coincidencias entre cadenas de palabras de largo uno, esto es debido a la limitada cantidad de palabras que presentan los mensajes.

Es el método de evaluación más usado en resúmenes utilizando redes sociales. Los jueces seleccionan manualmente cuales tweets creen que son los más importantes y que deberían estar en el resumen, este conjunto pasa a ser el resumen de referencia, luego se calcula ROUGE, con el resumen generado automáticamente Bian et al. [2015], Sharifi et al. [2010], Meng et al. [2012], Shou et al. [2013], Štajner et al. [2013], Yan et al. [2011].

El principal problema con utilizar ROUGE es que busca calces exactos entre las palabras, por lo cual su precisión baja si existen resúmenes que digan lo mismo, pero de forma distinta.

Método de la Pirámide Es un método de evaluación semi-automática, propuesto por Nenkova and Passonneau [2004]. Este método consiste en identificar un conjunto de elementos denominados unidades de contenido de resumen (SCU), la cuales son usadas para comparar la información entregada. Los SCU surgen de anotaciones que no son mayores a una cláusula en un conjunto de resúmenes. Para encontrar los SCU, primero se identifican sentencias similares y luego mediante una inspección más detallada encontrar las subpartes que estén más relacionadas. De esta forma las SCU que aparezcan en más resúmenes tendrán un peso mayor, siguiendo este proceso se construye una pirámide a partir de la anotación de SCU's en los resúmenes de referencia. En la cúspide de la pirámide estarán las SCU con mayor peso. Luego se anotan las SCU para el resumen generado y se comparan las pirámides de ambos conjuntos, para medir cuanta información es compartida entre ambos.

Divergencia de Kullback-Leibler Es una métrica, propuesta por Kullback and Leibler [1951], para medir las diferencias entre dos distribuciones de probabilidades. Está definida por la siguiente ecuación.

$$KL(x, y) = \sum_{\mathrm{i}} x(\mathrm{i}) \log \frac{x(\mathrm{i})}{y(\mathrm{i})} \tag{4.9}$$

Esta métrica debe interpretarse como, cuanta información se está perdiendo al representar la distribución x mediante la distribución y.

Permite evaluar un resumen sin la necesidad de utilizar un gold standard, ya que se

comparan los resúmenes contra el conjunto original de documentos, midiendo cuanta información se ha perdido.

Su principal problema es que no es simétrica y tampoco cumple la desigualdad triangular.

Divergencia de Jensen Shannon De acuerdo a lo propuesto por Dagan et al. [1997], la divergencia de Jensen-Shannon es una medida para cuantificar la divergencia entre dos distribuciones de probabilidades. Tiene la ventaja que es simétrica y finita , es utilizada como métrica en el trabajo deNenkova and Passonneau [2004] y de Mackie et al. [2014]. Está definida de la siguiente forma

$$JSD(x, y) = \frac{1}{2}D(x|M) + \frac{1}{2}D(y|M) \tag{4.10}$$

Donde $M = \frac{1}{2}(x + y)$ y D() es la divergencia de Kullback-Leibler.

Fraction of Topic Words Mide el cociente entre la cantidad de *topic words* que existen en el conjunto original de documentos, versus las presentes en el resumen. *Topic words* corresponden a unidades textuales (palabras, bigramas, etc), que son altamente descriptivas sobre el documento original, es decir son términos que tienen una importancia estadística mayor para un documento especifico que sobre todo el conjunto de palabras, fue desarrollado por Lin and Hovy [2000].

La principal ventaja de esta medida es que puede ser usada con o sin un gold standard, ya que se puede utilizar para comparar las distribuciones de probabilidades de las palabras entre el resumen automático y uno humano o puede ser usada para comparar la distribución entre el resumen y los documentos originales Louis and Nenkova [2013]. Además, acorde al trabajo de Mackie et al. [2014], se ha demostrado una buena correlación entre esta métrica y evaluaciones hechas con jueces humanos en métodos que utilizan datos de Twitter como conjunto de entrada.

4.1.2. Evaluación extrinsica

Estos tipos de evaluaciones consisten en ver como el resumen ayuda a la resolución especifica de una tarea. Pueden usarse distintas tareas, entre las tareas que se pueden utilizar para esta evaluación tenemos *categorización de documentos* y *preguntas y respuestas*. La primera consiste en utilizar el resumen para poder asignar una categoría al documento original, se comparan las diferencias con respecto a utilizar el documento original para categorizar. La segunda consiste en realizar preguntas de comprensión sobre el evento a un conjunto de usuarios.

Este tipo de evaluación es menos usada, tiene un costo más alto, ya que requiere planificar una tarea específica para los usuarios, la cual debe ser acotada y claramente especificada, además de reunir un número importante de voluntarios para que realicen la tarea diseñada.

Este enfoque es abordado por Marcus et al. [2011]. En el cual buscan medir si su sistema propuesto facilita la comprensión o no de un evento, para esto les piden a los voluntarios que respondan una serie de preguntas relacionadas al evento, como por ejemplo: determinar

donde y cuando sucedió un sismo. Otra tarea que le asignaron, a los voluntarios, fue que en una cantidad limitada de tiempo debían explorar el resumen, para luego ellos escribir su propio reporte del evento.

4.1.3. Discusión

Como ya se explicó anteriormente la evaluación de cualquier tipo de resumen es un problema complejo, existen diversas respuestas correctas y válidas para un mismo documento. Qué tan correcta sea una respuesta depende de diversos factores como: quien leerá el resumen, para qué lo utilizará, en que momento del tiempo lo utilizara. Por lo cual la principal dificultad al momento de evaluar un resumen es poder determinar qué cosas son importantes y cuáles no. Es muy difícil desarrollar una métrica que considere este aspecto para cualquier resumen. Por lo cual las métricas presentadas poseen algunas deficiencias.

En el caso de una evaluación extrínseca, esta permite evaluar distintos aspectos del resumen y como los usuarios interactúan con el mismo. El principal problema de esta metodología es que es muy costosa, ya que implica definir una tarea especifica y reunir una cantidad suficiente de usuarios que la ejecuten. Además, la ejecución de los usuarios es propensa a errores, por lo cual se deben agregar mecanismos de validación.

En el caso de la evaluación basada en calidad de texto, presenta un problema similar a la evaluación extrínseca, ya que se necesitan usuarios externos que evalúen los resúmenes y asignes una nota al resumen según los criterios que se estén evaluando. Además, si la selección de criterios no es la adecuada puede producir resultados incorrectos. Por ejemplo, un grupo de resúmenes podrían estar gramaticalmente bien escritos, poseer coherencia y estructura en sus frases y no ser redundante y aun así no ser una representación incorrecta del evento particular. En otro caso, si se seleccionan solo mensajes de spam como parte del resumen de un evento, estos cumplirán con las propiedades anteriores pero no representaran correctamente el evento.

En el caso de las medidas basadas en contenido, algunas dependen de un conjunto de resúmenes generados por humanos para comparar como es el caso de ROUGE y el método de la pirámide. Esto tiene como desventaja que es muy costoso, ya que hay que juntar a un grupo de personas para que generen los resúmenes, esto también puede introducir sesgos en la forma en como los usuarios generan el resumen. Otro problema con la construcción de este conjunto es que se asume de forma inmediata que la calidad es la indicada, esto no necesariamente cierto, ya que los usuarios pueden asignarle distinta importancia a distintos tópicos dentro del resumen, o se pueden ver influenciados por otros efectos externos. Esto se puede mejorar utilizando usuarios que sean expertos en el área, para la generación de los resúmenes, pero esto no siempre es posible.

Otro problema que se debe considerar que es transversal a los distintos tipos de evaluación es: En los distintos trabajos presentados, se utilizan datasets distintos con eventos de distinta naturaleza (desastres naturales, deportes, terrorismo, etc). Por ejemplo, los trabajos realizados por Bian et al. [2015], Metzler et al. [2012], Schinas et al. [2016] utilizan como fuentes de datos eventos seleccionados al azar a partir de conjunto mayor de mensajes, esto produce que los eventos presentes en dichos conjuntos sean distintos e incomparables. Los autores de los distintos trabajos utilizan esta metodología para extraer mensajes debido a

que no existen un gold standard único contra el cual comparar para resúmenes usando redes sociales. Esto tiene como principal desventaja que hace muy difícil comparar dos métodos distintos. Primero, porque los conjuntos de mensajes utilizados por ambos serán totalmente distintos. Segundo, si existen usuarios involucrados estos generaran resúmenes distintos, aunque se utilizaran los mismo mensajes.

Un problema adicional detectado en los trabajos relacionados, fue que algunos trabajos le asigna un puntaje a los mensajes seleccionados en base a distintos criterios, como es el caso del trabajo de Bian et al. [2015], que asigna puntajes según *cobertura*, *importancia* y *diversidad*, pero en la evaluación no se mide el impacto de estos criterios en el resultado final de los mensajes seleccionados.

De acuerdo a lo presentado en este capítulo, podemos responder la pregunta de investigación número uno presentada en el capítulo 1. Si bien existen distintas métricas de evaluación para resúmenes automáticos la gran mayoría de los trabajos enfocados en redes sociales utilizan ROUGE como métrica de evaluación. Aunque los trabajos utilizan esta métrica como base para determinar la calidad de un resumen, las evaluaciones realizadas entre los trabajos no son comparables. La dificultad para realizar dicha comparación radica en que ROUGE se basa en comparar el resumen generado con uno generado de forma manual, esto dificulta las comparaciones, porque los resúmenes producidos de forma manual casi siempre serán distintos, dado que los usuarios que los generan son distintos.

En está investigación el enfoque utilizado corresponde a una evaluación intrínseca, basada en medidas de similitud de contenido. La métricas utilizadas para este trabajo fueron: *Similitud coseno, indice de Jaccard, divergencia de Kullback-Leibler, divergencia de Jenseen-Shannon, Topic words*. Se escogió este enfoque, ya que no requiere de un conjunto de resúmenes generados de forma manual, esto permite evaluar los resúmenes de forma automática. Para todas las métricas se compararon el conjunto inicial de mensajes versus el conjunto de resúmenes generados.

Capítulo 5

Descripción Experimental

El objetivo principal de este trabajo es comparar distintos métodos para la generación de resúmenes automáticos. Para poder realizar esto primero se debe contar con un dataset que contenga eventos noticiosos con distintas características, para comprobar que los métodos no tengan un sesgo hacia un tipo de evento particular. Además, contar con distintos métodos para poder comparar su desempeño. En este capitulo se presenta el conjunto de datos utilizado correspondiente a cuatro eventos distintos, además, se explica la metodología con la cual fueron detectados los eventos y recolectados los mensajes. También, se presentan los métodos seleccionados para la comparación. Finalmente, se explica los experimentos realizados, consistente en un estudio de casos en el cual se usaron métricas automáticas de evaluación y un estudio de los tópicos presentes en los resúmenes.

5.1. Recolección de eventos

Para recolectar y detectar eventos noticiosos en Twitter, se utilizó la metodología propuesta en Kalyanam et al. [2016b]. Esta metodología, consiste en detectar pares de términos que tienen un alto nivel de actividad en un corto tiempo. A continuación, se especifican los pasos de la metodología.

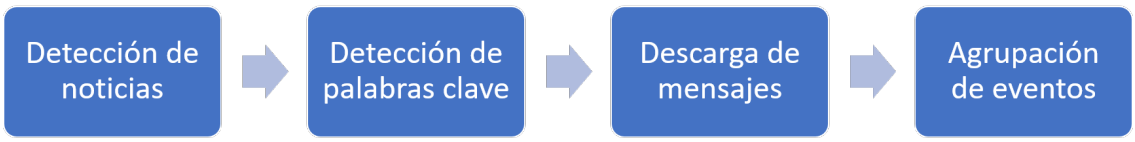

Figura 5.1: Diagrama sobre flujo de recolección de eventos

1. **Detección de noticias** Primero se obtienen titulares de noticias desde un conjunto de cuentas verificadas que pertenecen a distintos medios de comunicación como CNN, BBC, USAToday, entre otros. Estos titulares son obtenidos cada una hora.

2. **Detección de palabras claves** El objetivo de esta etapa es detectar tuplas de palabras

claves, que describan de forma coherente un evento. Para esto se optó por una estrategia iterativa, dado un conjunto de titulares S, extraído en la parte anterior se procede a: Primero por cada par de titulares se calcula su intersección si el tamaño de dicha intersección es mayor a cierto umbral, el par es considerado. Segundo, si ese nuevo conjunto tiene un índice de Jaccard mayor a otro umbral con respecto a los conjuntos creados en iteraciones anteriores, entonces se agrega este par a dicho conjunto, para el caso contrario, se crea un nuevo conjunto independiente. Los componentes de la tupla pueden ser tanto palabras como entidades nombradas, esto porque muchos mensajes pueden referirse al mismo elemento, pero con términos distintos o compuestos, por ejemplo, la palabra "Trump" y el término "Donald Trump" se refieren a lo mismo, por lo cual para agrupar mejor los términos claves de los eventos, se optó por detectar entidades nombradas en texto del mensaje y utilizarlos también como elementos pertenecientes a la tupla. El reconocimiento de entidades nombradas fue agregado en una segunda versión de este algoritmo. El puntaje de cada tupla corresponde al puntaje promedio de cada uno de los términos que la componen y el puntaje de cada término es la cantidad de veces que fue agregado a una tupla. tercero y final, se seleccionan los 2 elementos con el puntaje más alto.

Al momento de detectar los términos para conformar las tuplas, puede suceder que se unan términos que aparecen en eventos distintos y que no aportan información, por lo general corresponden a términos que son comunes a través de los titulares de noticias, por ejemplo, "watch", "live", "update", entre otros.

Para solucionar este problema se utilizó una versión modifica de *tf-idf*, denominado *maxtf-idf*. En tf-idf, se busca determinar qué tan rara es una palabra para la colección y que tan frecuente es en un documento. Por otra parte, con *maxtf-idf*, se busca asignar un peso mayor a los terminos que son más frecuentes en cualquier documento, con respecto a su frecuencia en el documento actual.

maxtf-idf está definido de la siguiente forma:

$$maxtf(t, \mathrm{d}, D) = 0{,}5 + \frac{0{,}5 + max\{f(t, \mathrm{d}') : \mathrm{d}' \in D\}}{max\{f(w, \mathrm{d}) : w \in \mathrm{d}\}} \tag{5.1}$$

Donde t es una palabra, d es un documento y D es la colección completa de documentos. Para este caso en particular se estableció t como una de las palabras claves, d como el conjunto de palabras claves de una hora especifica en un día en particular, y D como el conjunto de documento de ese mismo día.

Luego de haber identificado estos términos, se debe desconectar los conjuntos unidos mediante estos términos. Los conjuntos se separan escogiendo el termino con el mayor puntaje *maxtf-idf*, este proceso se repite hasta que los eventos no pueden ser separados más. Finalmente, las palabras encontradas por este proceso son agregadas a la lista de stopwords para no ser consideradas más en la recolección de eventos.

3. Descarga de mensajes Luego utilizando la API de desarrollador de Twitter [1], se utilizan los términos encontrados en la parte anterior como palabras claves de búsqueda

[1] https://developer.twitter.com/en.html

y a partir de eso se descargan los tweets que contengan dichos términos. Finalmente, un evento noticioso e está caracterizado como la tupla *(keyword$_1$,..,f,tweet$_1$,..,tweet$_m$)*, donde f es la fecha de descarga de los mensajes.

4. **Agrupación de eventos similares** Debido a que los eventos en redes sociales van cambiando con el tiempo, pueden existir eventos detectados en la parte anterior que tengan fechas distintas, pero están relacionados con otro evento. Para solucionar esto se unen los conjuntos que al menos tengan un termino en común. Este proceso se realiza tomando los eventos detectados en las ultimas 24 horas.

5.2. Eventos seleccionados

Para la evaluación de los distintos métodos de resúmenes, se seleccionaron distintos eventos del mundo real. Los eventos fueron extraídos desde Twitter utilizando la metodología de recolección de eventos anteriormente descrita.

Los eventos seleccionados son de diversos tipos con distintas duraciones, algunos duran solo un par de días (corta duración), mientras otros se extienden por una semana (duración media); y otros por un mes o más (larga duración); distintas entidades involucradas, por ejemplo, países o personas naturales; distintas ubicaciones geográficas, pero todos tienen en común que causaron un impacto considerable dentro de la red social. Se seleccionaron eventos de distinto tipo para evitar que hubiese un sesgo al momento de evaluar los distintos métodos, ya que puede ser que alguno de los métodos este optimizado para eventos con características especiales, como una duración temporal corta o que fuesen eventos programados.

A continuación, se detallan los eventos seleccionados explicando sus principales características, además de análisis exploratorio hecho para cada evento:

Huracán Irma

Corresponde a un evento de duración media que va desde el día 30 de agosto hasta el 10 de septiembre. Durante este periodo la tormenta tropical y posterior huracán llamado Irma, paso por todo el caribe y por el estado de Florida en Estados Unidos Donegan [2017. Fecha Acceso: 2017-11-09], Levenson [2017. Fecha Acceso: 2017-11-09].

Fue catalogado como el huracán más fuerte y destructivo en el Atlántico desde el 2005. Causando enormes daños y muchas muertes en Islas Vírgenes, Barbuda, Puerto Rico, Cuba, Haití, Estados Unidos entre otros.

Los mensajes corresponden al paso del huracán Irma por el estado de Florida en Estados Unidos, incluyendo la evacuación del estado, inundaciones producidas, saqueos, información otorgada por las autoridades, entre otros.

Juicio de Oscar Pistorius

Este es evento un evento de larga duración, que corresponde al juicio al cual fue sometido el atleta sudafricano Oscar Pistorius por el asesinato de su novia Reeva Steenkamp. La duración del juicio fue desde el 3 de marzo del 2014 hasta el 12 de octubre del 2014, en el

cual se encontró culpable al atleta Phipps [2014. Fecha Acceso: 2017-11-09], Fihlani [2015. Fecha Acceso: 2017-11-09].

En este caso en particular los tweets extraídos van desde el 15 de febrero del 2013, el día que sucedió el crimen, hasta el 8 de agosto del 2014.

En el juicio se registraron diversos testimonios de testigos, interrogatorio a Oscar Pistorius, reacciones de los medios de comunicación y las opiniones que se generaron en la red social. En este caso la recolección de eventos no logro extraer la parte final del juicio con su veredicto.

Terremoto en Nepal

El 25 de abril del 2015, ocurrió un terremoto grado 7.8 en Nepal. Considerado el mayor terremoto registrado en el país. Provocando la muerte de más de 9.000 personas e hiriendo a otras 22.000, además de severos daños a las construcciones e infraestructura del país Yanrong [2015. Fecha Acceso: 2017-11-09].

Los tweets extraídos corresponden al 25 de abril del 2015, incluyendo la cobertura internacional del terremoto, el conteo de víctimas fatales y los daños sufridos en las distintas ciudades del país.

Ataque terrorista en Libia

El día 27 de abril del 2015, un grupo de 5 hombres armados con afiliaciones al estado islámico, irrumpen en el hotel Corinthia en Tripoli, la capital de Libia, disparando contra los huespedes y trabajadores. Este ataque dejo 9 muertos entre un ciudadano estadounidense y uno francés, además de todos los terroristas Jawad [2015. Fecha Acceso: 2017-11-09].

Los mensajes descargados corresponde al 27 de abril del 2015, abarcando todo el evento, desde el primer ataque hasta la resolución del conflicto. Incluyendo reportes de autoridades e información minuto a minuto de la crisis.

En la tabla 5.1 se muestra un resumen de las características principales de los mensajes extraídos para cada evento.

	Huracan Irma	Juicio Oscar Pistorius	Terremoto en Nepal	Ataque en Libia
Tweets	697.795	113.215	522.804	28.640
Originales	685.372	85.386	154.112	16.037
Retweets	9.065	26.272	363.805	12.292
Replies	3.358	1.557	4.887	311
Imágenes Únicas	14.702	911	10.083	1.389
Urls	698.191	66.853	457.114	22.939

Tabla 5.1: Información de cada evento

5.3. Métodos seleccionados

A continuación, se explican los distintos métodos seleccionados, para generar resúmenes de los eventos seleccionados. Los métodos seleccionados representan las distintas técnicas descritas en el capítulo 3, además de representar distintos grados de complejidad.

Mgraph

El primer método seleccionado fue MGraph, propuesto por Schinas et al. [2016]. Este método se encuentra explicado en el capítulo 3. La implementación de este trabajo fue provista por los autores. Para generar los descriptores de las imágenes e índices utilizados por el método se utilizo la implementación del trabajo realizado por Spyromitros-Xioufis et al. [2014].

Phrase Reinforcement

El segundo método seleccionado fue una variante de Phrase Reinforcement Algorithm propuesto por Sharifi et al. [2010], el cual fue explicado con en el capítulo 3. La única diferencia con la versión original del método, es que antes de aplicar el algoritmo, se aplico K-means sobre el conjunto de mensajes, para poder detectar los subeventos del evento, posteriormente se aplico Phrase Reinforcement sobre cada clúster, utilizando como nodo raíz el termino más representativo de cada clúster. Para aplicar K-means se calcularon los vectores tf-idf de cada mensaje, descartando mensajes que tuvieran más de 2 urls; más de 3 hashtags o más de 2 menciones de usuarios. Luego se aplico K-Means sobre el conjunto restante, utilizando como métrica, la distancia euclidiana. El número de clusters para cada evento fue determinado de forma empírica.

K-means + Centroid

Este es el método más simple, consiste en primero calcular los vectores tf-idf sobre el conjunto de tweets de un evento, considerando solo el texto del mensaje, antes de calcular estos vectores se removieron stop words y urls, ademas se descartaron los mensajes que tengan más de 2 urls; más de 3 hashtags o más de 2 menciones de usuarios, se escogió esta heurística, porque los tweets que cumplan con alguna de estas condiciones poseen muy poca información para ser considerados relevantes. Luego sobre estos vectores se aplica k-means, utilizando distancia euclidiana como medida de distancia. El número de clúster fue determinado empíricamente para cada evento en particular. Luego una vez formados los clústeres, se busca entre todos los tweets de un mismo clúster, el tweet más cercano al centroide de dicho clúster, en este caso ahora se utiliza distancia coseno para determinar el tweet más cercano, este mensaje es seleccionado como representante del clúster y pasara a formar parte del resumen, finalmente, se selecciona un mensaje por cada cluster.

Word Embbedings + Clustering

Este método fue realizado en conjunto con Mauricio Quezada en el marco de su tesis doctoral. Para desarrollar este método se probaron distintas representaciones vectoriales y distintas técnicas de clustering:

1. **Generación de Documentos:** Para reducir la redundancia entre los mensajes, estos son agrupados por la urls que contienen, es decir todos los tweets que contengan la misma url, pasan a formar un solo conjunto. Esto porque los mensajes que comparten la misma url, suelen estar asociados al mismo tema, incluso pueden ser simplemente el mismo mensaje publicado muchas veces. En esta etapa además se descartan todos los mensajes que contengan más de 3 hashtags o más de 2 urls. Luego de cada conjunto de tweets, ya que dichos tweets poseen muy poca información textual que pueda ser relevante. El conjunto de tweets agrupados se denomina documentos, luego por cada documento se extrae un representante. El representante de cada documento es el mensaje que haya tenido más retweets, en caso que no hayan retweets, se considera primer tweet original publicado.

2. **Representación Vectorial de Documentos** Ahora por cada representante se calcula una representación vectorial. Las representaciones utilizadas son tf-idf, fastText y glove. Para fasttext, se entrenó el modelo utilizando todos los tweets disponibles del corpus. En el caso de Glove se utilizo un modelo ya entrenado por los autores del método [2], el cual fue entrenado sobre un conjunto de dos mil millones de tweets.

3. **Clustering** Una vez obtenido los vectores representantes de cada documento, se aplica clustering para poder detectar los sub-tópicos más relevantes. La cantidad de clusters fue determinada empíricamente para cada evento en particular. Se probaron dos variantes de clustering, K-means y aglomerativo, en el caso del clustering aglomerativo se probo con complete y average como métodos de enlace. Para K-means se utilizo la distancia euclideana como medida de distancia entre los elementos.

4. **Ranking por Impacto** Un evento tiene tópicos que son más importantes que otros, que atraen más atención en las redes sociales. Con el fin de capturar esta característica, se realiza un ranking de los clústers para medir su impacto. Este ranking esta basado en el enfoque propuesto por Kalyanam et al. [2016b], para clasificar eventos de alto impacto. Para esto primero obtenemos y ordenamos todas las fechas de creación de los mensajes, luego se calcula la diferencia de tiempo entre 2 tweets consecutivos, esto se repite para todos los mensajes del evento. A continuación se calcula un histograma con las diferencias obtenidas. Finalmente se ordenan los histogramas en orden decreciente del tamaño del primer compartimiento del histograma. Este ranking favorece a clusters que sean grandes y que hayan generado un gran volumen de datos en un poco tiempo, es decir, ubicara en posiciones más altas los tópicos que concentraron una mayor discusión en la red social.

5. **Selección de Representantes** Finalmente, para mostrar los mensajes al usuario se seleccionan k mensajes, donde k es el número de tópicos seleccionados. Para seleccionar estos mensajes, primero se calcula el centroide de cada clúster, luego se selecciona el mensaje más cercano a dicho centroide, para formar parte del resumen final.

[2] https://nlp.stanford.edu/projects/glove/

5.4. Análisis experimental

Una vez generados los resúmenes para cada uno de los eventos utilizando los algoritmos anteriormente mencionados, se deben establecer criterios de comparación entre ellos con el fin de determinar su calidad. En esta sección se presentan dos tipos de análisis uno enfocado en comparar los resúmenes con conjuntos de datos que también describen el evento. El segundo análisis consiste en evaluar la presencia de los distintos tópicos existentes del evento en el resumen.

5.4.1. Análisis exploratorio

En este primer análisis se comparan los resúmenes de los distintos métodos contra dos conjuntos, el primero son noticias de medios tradicionales y el segundo es el conjunto original de todos los mensajes de un evento. Para ambos casos se calcularon métricas automáticas de cobertura y diversidad.

Comparación con medios tradicionales

Debido a la ausencia de un conjunto de resúmes de prueba (ground truth) para los eventos seleccionados y en general para resúmenes utilizando redes sociales, se buscó un reemplazo para esto y se optó usar las noticias publicadas en los medios tradicionales, esto porque los medios tradicionales realizan una cobertura general y objetiva de los ocurrido en una red social. En el caso de eventos más largos como el juicio de Oscar Pistorius, se buscaron resúmenes o líneas de tiempo publicadas por medios tradicionales. Las noticias fueron seleccionadas manualmente. Esta evaluación tiene dos objetivos, el primero ver si lo comentado en redes sociales es similar a lo publicado en los medios tradicionales y el segundo corresponde a ver si se pueden utilizar los medios tradicionales como un ground truth para la evaluación cuando no existe uno. Las métricas de comparación utilizadas fueron las siguientes:

Índice de Jaccard Se calcularon distintas versiones. Primero se calculó entre cada medio de referencia y el conjunto total de tweets, considerando solo los términos más frecuentes, hasta 10 términos; Luego entre la concatenación de todos los medios de referencia y el conjunto total de tweets, considerando los términos más frecuentes con un tope de 10 términos, entre cada medio de referencia y cada resumen generado. Para cada una de las distintas versiones se eliminaron los stop words y se aplicó stemming a cada uno de los términos, tanto en los resúmenes, como en los medios de referencia y en el conjunto de tweets.

Divergencia de Jensen Shannon Se calculó la distribución de probabilidades de las palabras tanto del resumen como del medio de referencia. Además, se calculó la divergencia entre todos los tweets del evento y la concatenación de todos los medios de referencia. Se eliminaron stop words y se aplicó stemming a las palabras.

Evaluación de cobertura

Un aspecto importante que se debe considerar en cualquier resumen, es que tenga una alta *cobertura*. Esto es que el resumen generado abarque todos los contenidos importantes del conjunto inicial de documentos. El fin principal de estos resúmenes es informar sobre

el evento, por lo cual se busca que el resumen sea una buena representación del conjunto original. Para esta evaluación se utilizó el enfoque propuesto por Nenkova and Passonneau [2004], que consiste en evaluar el resumen contra el conjunto original de documentos, en este caso son todos los tweets asociados a un evento. Se utilizaron las mismas métricas que en Nenkova and Passonneau [2004].

Divergencia de Kullback-Leibler Es calculada entre la distribución del input y el resumen

Divergencia de Jensen-Shannon Calculada entre la distribución del input y el resumen.

Similitud coseno Considera la similitud coseno entre todos los términos del input y todos los términos del resumen.

Porcentaje topic tokens Proporción de términos presentes en el resumen que son topic words en el conjunto original.

Fracción topic words Proporción de términos que son topic words en el conjunto de entrada que aparecen en el resumen.

Calce de topic words Similitud coseno entre todos los términos del resumen y las topic words presentes en el conjunto de tweets

Evaluación de diversidad

Otro aspecto importante al momento de generar un resumen es evitar repetir información, ya que no aporta a la comprensión del evento y además implica que no se está mostrando información que es potencialmente relevante.

La evaluación de cobertura no es capaz de medir esto, ya que solo se está comparando si los mensajes seleccionados son una buena representación del conjunto original de datos. Es por esto que se decidió realizar una evaluación solo para medir este aspecto.

Esta evaluación consiste en comparar el contenido textual de cada mensaje en el resumen, contra todos los demás mensajes seleccionados en el mismo resumen, para luego tomar el valor promedio de todos los mensajes. Las métricas utilizadas para esto son:

Índice de Jaccard Índice de Jaccard promedio entre todos los mensajes de un mismo resumen.

Similitud Coseno Similitud coseno promedio entre todos los mensajes del resumen.

Distancia de Levenshtein Distancia de Levenshtein promedio entre todos los mensajes. La distancia de Levenshtein corresponde al número mínimo de operaciones necesarias para convertir una cadena de texto en otra.

5.4.2. Análisis cualitativo

Si bien la evaluación automática entrega información muy importante sobre la calidad del resumen, posee algunas deficiencias. Por ejemplo, existen mensajes que comunican lo mismo, pero de distinta forma. Entonces si solo se consideran sus diferencias textuales, estas dirán que son distintos, pero en realidad es información repetida que no aporta a la comprensión del evento. Por otra parte, la evaluación de diversidad no considera si el contenido de los mensajes es relevante para el usuario. Podría suceder que un método seleccionara puros mensajes spam, que usaran términos distintos y las métricas utilizadas indicarían que tienen una alta diversidad.

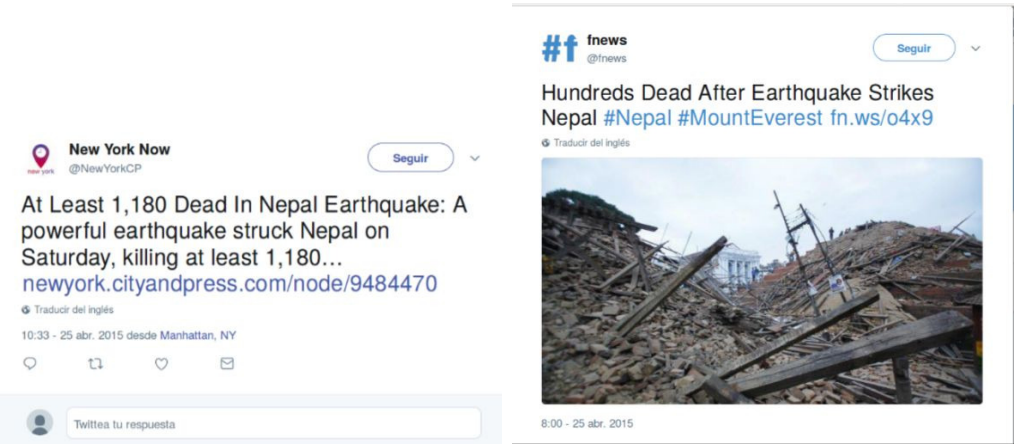

Figura 5.2: Tweets sobre el terremoto en Nepal

Los dos mensajes mostrados en la figura 5.2 tienen muy pocos términos en común, por lo cual todas las métricas indican que son distintos, pero es fácil comprender que ambos mensajes hablan de lo mismo, en este caso sobre el número de muertos.

Es por las situaciones descritas anteriormente que es necesario complementar las métricas automáticas con un un análisis cualitativo más detallado sobre los tópicos presentes en los mensajes.

5.4.3. Evaluación por tópicos

Esta evaluación tiene como fin validar que los resúmenes generados efectivamente sean una buena representación del evento. Que cubran todos los aspectos relevantes del evento repitiendo la menor cantidad de información posible.

Un tópico en un evento corresponde algún hecho especifico de alta importancia. Por ejemplo, un tópico para el evento del terremoto en Nepal es la coordinación de ayuda internacional, en el caso del ataque terrorista en Libia es la proclamación del atentado por parte del estado islámico.

Para esta evaluación se identificaron manualmente una lista de tópicos para cada evento. Luego de forma manual se inspeccionaron los resúmenes y se lleva un conteo de cuantos tópicos aparecen en el resumen y cuantas veces aparece cada uno.

Para identificar la lista de tópicos de cada evento se revisaron distintas publicaciones sobre los eventos en medios tradicionales y se anotaron los tópicos presentes, luego se revisó manualmente si existían tweets que correspondieran a dichos tópicos, además al momento de revisar el conjunto de tweets también se agregaron nuevos tópicos que no aparecieron originalmente en los medios tradicionales.

Para esta evaluación se consideraron como relevantes, aquellos mensajes que hablen sobre algo en específico del evento. Por ejemplo, la figura 5.3, está relacionada con el evento en particular, pero no aporta ninguna información especial, para comprender el evento.

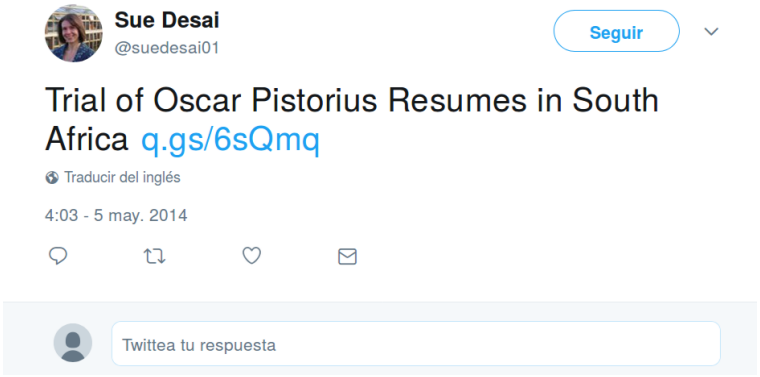

Figura 5.3: Ejemplo de tweet no informativo

Capítulo 6

Análisis de los resultados

En este capítulo se presentan los resultados obtenidos para cada uno de los experimentos, mencionados en el capítulo anterior, y que tratan de la aplicación de los cuatro métodos selecciones en este trabajo. Primero se presentan los resultados obtenidos al comparar los resúmenes automáticos con los medios tradicionales de noticias. Luego, se presentan los resultados obtenidos para la evaluación enfocada en cobertura. A continuación, se presentan los resultados de la evaluación enfocada en diversidad y finalmente se muestran los resultados obtenidos para el análisis cualitativo de tópicos. Para cada una de las evaluaciones se presentan los resultados, incluyendo un análisis de los valores obtenidos, y buscando las posibles causas que explique las diferencias obtenidas entre los métodos.

Para la aplicación de los diferentes métodos, se tiene una particularidad a considerar con el método word emmbedings + clustering, ya que trata de un método que es el resultado de otro trabajo conjunto con está tesis, luego, resulta necesario probar diferentes variaciones del mismo, para validar su funcionamiento y grado de confiabilidad. Las variantes seleccionadas son:

1. **K-means FastText** K-means sobre vectores generados con fasttext.

2. **K-means Glove** K-means sobre vectores generados con Glove.

3. **Aglomerativo 1** Clustering aglomerativo utilizando distancia euclidiana como medida de similitud entre los vectores, método complete como distancia entre clusters y representación vectorial generada con Glove.

4. **Aglomerativo 2** Clustering aglomerativo utilizando distancia coseno, método complete como distancia entre clusters y representación vectorial generada con Fasttext.

5. **Aglomerativo 3** Clustering aglomerativo utilizando distancia euclidiana, método complete como distancia entre clusters y representación vectorial generada con tf-idf.

En cada uno de los métodos de la lista, se seleccionaron como representantes, los centroides de cada clúster.

6.1. Exploración de los resultados

A continuación, se presentan los resultados obtenidos para cada uno de los experimentos realizados. Considerando las métricas descritas en el capítulo anterior.

6.1.1. Comparación con los medio tradicionales

Está comparación consistió en generar dos conjuntos de palabras para el mismo evento, uno derivado desde los medios tradicionales y otro obtenido a partir de los resúmenes generados. Para los análisis realizados se utilizaron como referencia tres documentos extraídos de medios de noticias internacionales, como BBC o CNN. Además los valores mostrados corresponden al promedio de evaluar cada documento por separado.

Resultados

En las figuras 6.1, 6.2 y 6.3 se muestran los valores obtenidos para las métricas.

	Huracán Irma	Juicio Oscar Pistorius	Terremoto en Nepal	Ataque en Libia
Jensen-Shannon	0.391	0.443	0.489	0.520
Índice de Jaccard	0.03	0.111	0.111	0.176
Kullback-Leibler	2.190	2.870	2.212	2.830

Tabla 6.1: Resultados considerando la concatenación de todas las líneas de tiempo y el conjunto completo de tweets

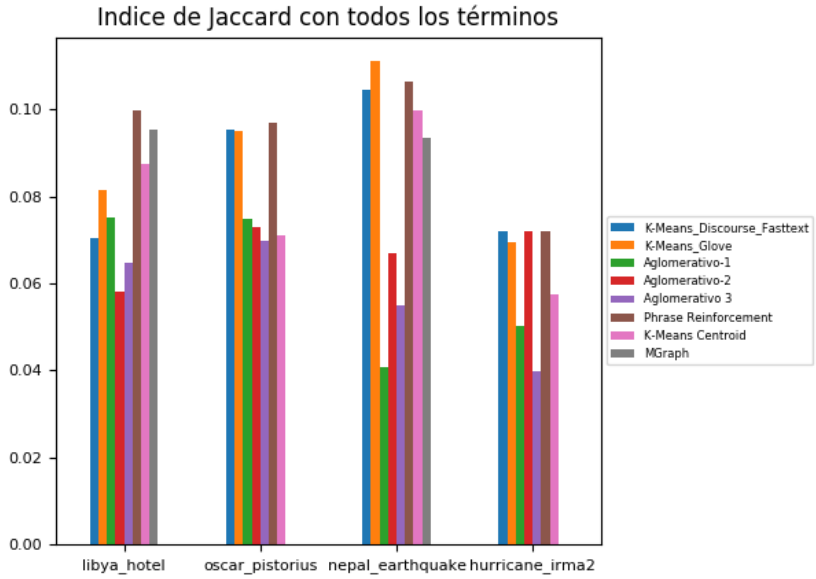

Figura 6.1: Resultados para índice de Jaccard considerando todos los términos

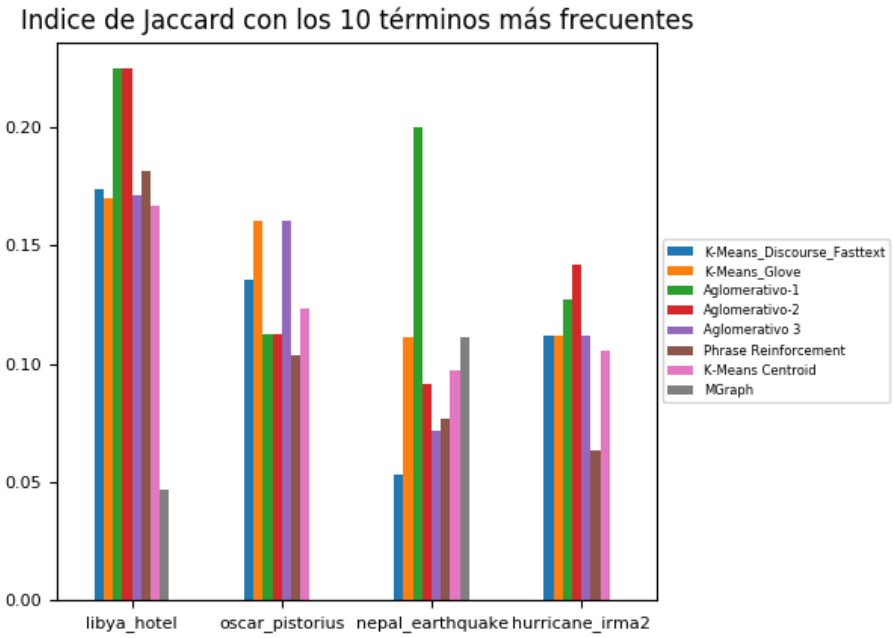

Figura 6.2: Resultados para índice de Jaccard considerando los 15 términos más populares por evento

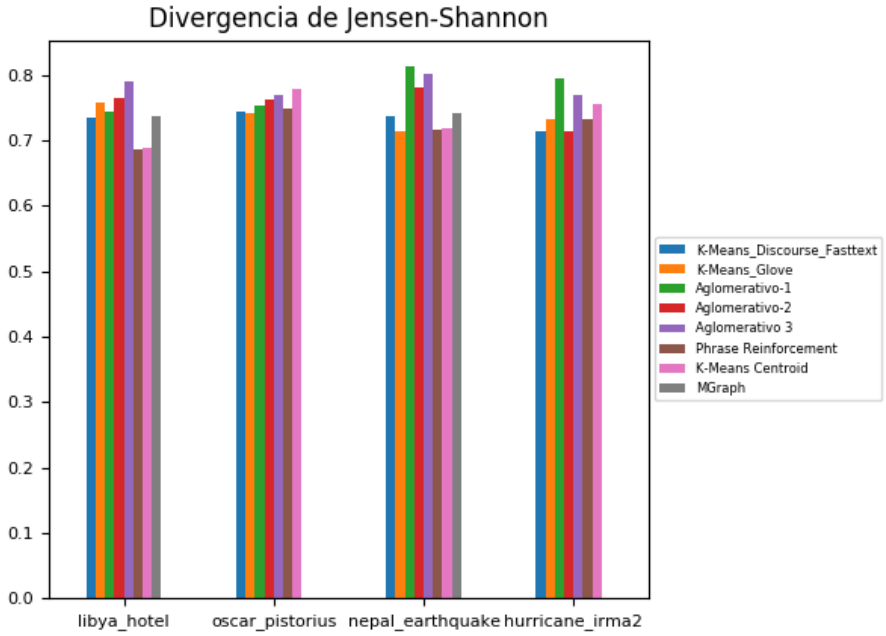

Figura 6.3: Resultados para Divergencia de Jensen-Shannon

Análisis

Como se observa en los resultados de la tabla 6.1 los resúmenes generados a partir de información de redes sociales no utilizan los mismos términos que los resúmenes obtenidos desde los medios tradicionales. Así mismo, al observar los valores del índice de Jaccard para cada uno de los eventos y cada uno de los métodos, se deduce que los términos utilizados por cada conjunto son distintos. Por otra otro lado, en los gráficos de las figuras 6.1 y 6.2 se aprecia que para todos los métodos y en todos los eventos, el porcentaje de términos similares varia entre el 10 % y 20 %. Esto nos permite establecer que ambos conjuntos no son comparables, siendo la consecuencia de ello que la información obtenida desde uno de ellos, difiere por sobre un 80 % en la información que puede entregar el otro conjunto. A partir del gráfico de la figura 6.3, se llega a la misma conclusión, ya que para todos los eventos los valores de la divergencia de Jesen-Shannon están por sobre 0.6 (la divergencia de Jensen-Shannon toma valores entre 0 y 1).

Esto puede deberse a los siguientes factores:

- La metodología de recolección de eventos (resúmenes automáticos) no haya capturado todos los mensajes asociados a un evento, con lo cual hay hechos particulares que no aparecerán en los mensajes, pero si en los medios tradicionales. A modo de ejemplo, para los casos de los eventos asociados al juicio de Oscar Pistorius y el Huracán Irma, se tiene; para el juicio de Oscar Pistorius, no se recolectaron tweets en la fecha que se entregó el veredicto del caso, por lo cual esto no aparece en los mensajes. Lo mismo sucede con el huracán, en donde se recolectaron mensajes a partir de la fecha en la cual el huracán estaba llegando a Florida y no se consideró su paso por otros países.

- Otro efecto que se produce es derivado desde el nivel de ruido existen en los tweets. En las redes sociales se encuentra un gran número de mensajes que pueden ser considerados como spam, es decir, mensajes que utilizan algunos términos claves del evento pero su contenido es sobre cualquier otra cosa. Por ejemplo, en marzo del 2018, 25.000 mensajes diarios fueron reportados por los usuarios como spam.

- En las redes sociales no está definido un formato y estilo de redacción requerido para generar contenido, permitiendo un lenguaje cargado de emociones y que responde al estilo de cada usuario. Está situación no se presenta en los medios tradicionales, ya que estos están sujetos a ciertas reglas y formatos. Sin querer establecer cual es mejor, se puede asegurar que generan resultados distintos.

- El foco que se mantiene para un evento en las redes sociales, dice relación con un punto particular del evento y en el caso de los métodos tradicionales esto se amplia a todo los que rodea y explica el evento. Por ejemplo, en el ataque terrorista en libia, los medios tradicionales además de cubrir el hecho en particular explicaban la compleja situación política del país en ese momento, esto no fue discutido en redes sociales

- Los procesos que se siguen para disponibilizar el contenido son distintos. En el de redes sociales, el proceso es escuchar y opinar y en el caso de los medios tradicionales el proceso es recolectar, investigar, validar, editar y publicar.

Estos resultados permiten responder la segunda pregunta de investigación presentada en el capitulo 1, con respecto a si la información publicada en redes sociales es similar a lo publicado en medios periodísticos tradicionales. Se observa a partir de la información de la tabla 6.1 y de los gráficos 6.1 y 6.2 que entre ambas fuentes la información compartida es distinta. Esto debido a como se explico anteriormente los vocabularios utilizados por ambos son distintos y lo publicado en los medios tradicionales esta libre de spam. Esto además implica que los medios tradicionales de noticias no pueden ser utilizados para evaluar resúmenes generados a partir de información de redes sociales, porque los temas discutidos en cada uno son diferentes.

6.1.2. Evaluación de cobertura

En la evaluación de cobertura se busca comparar si el vocabulario presente en el resumen generado automáticamente es similar al vocabulario presente en el conjunto original de mensajes asociados a un evento. Esto para establecer si el resumen es una buena representación de lo discutido en redes sociales.

Resultados

A continuación, en las figuras 6.4, 6.5, 6.6, 6.7 y 6.8, se presentan los resultados obtenidos por los distintos métodos, para cada una de la métricas de cobertura en cada uno de los eventos seleccionados.

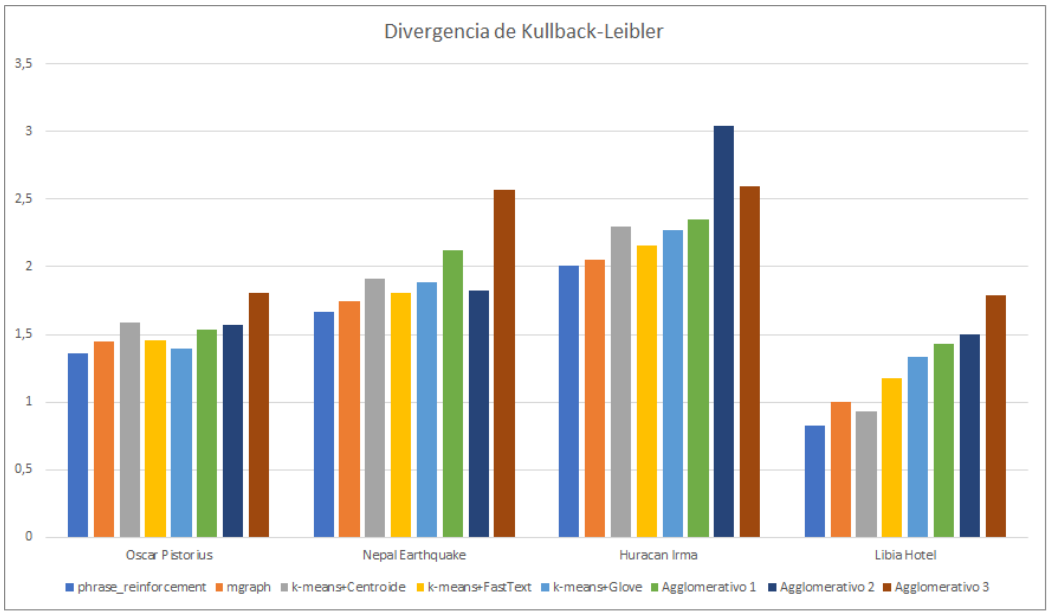

Figura 6.4: Resultados para Divergencia de Kullback-Leibler, entre el conjunto de mensajes y el resumen. Agrupados por evento

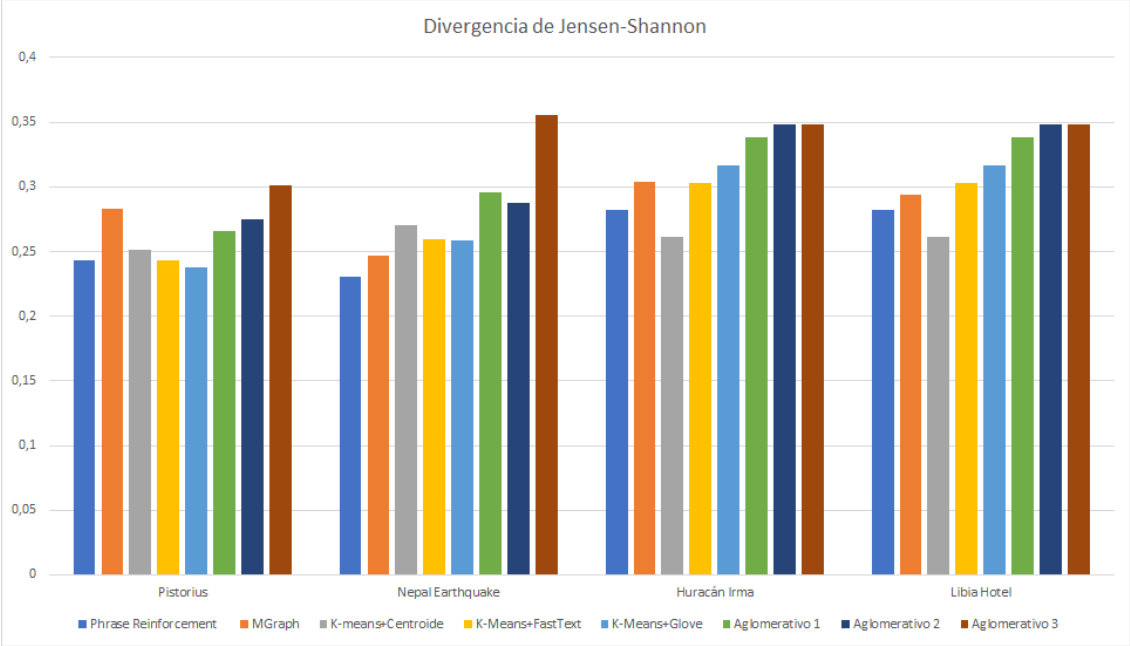

Figura 6.5: Resultados para Divergencia de Jensen-Shannon, agrupados por evento

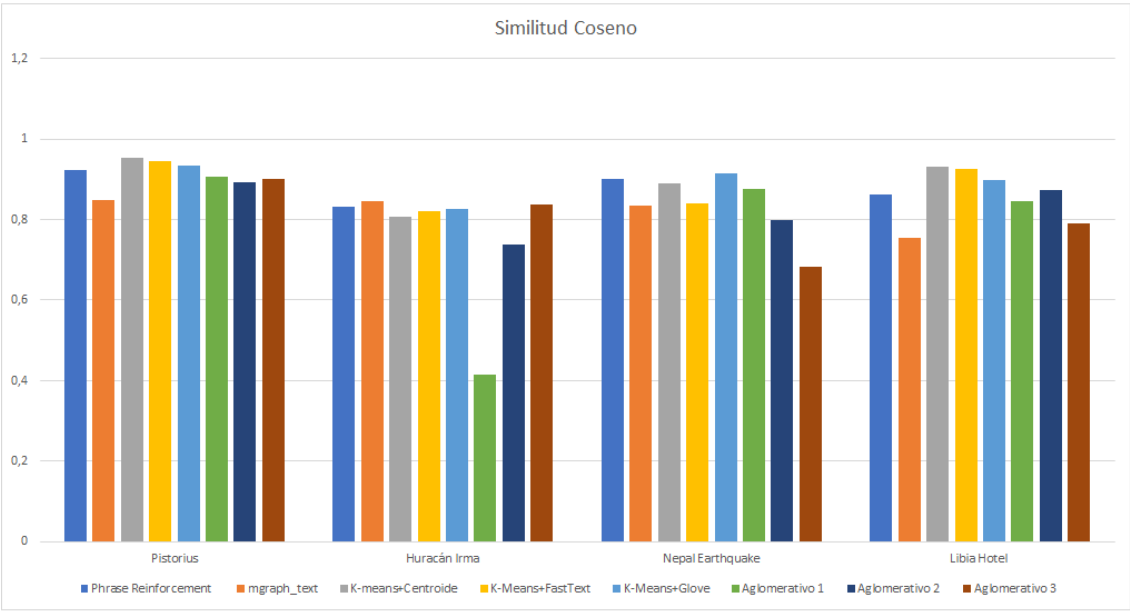

Figura 6.6: Resultados para Similitud Coseno, agrupados por evento

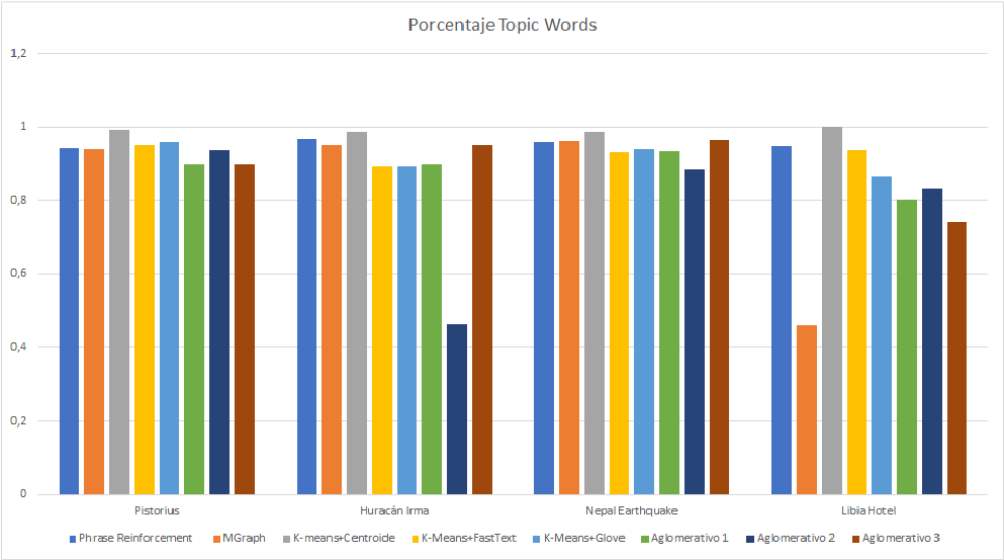

Figura 6.7: Resultados para Porcentaje de Topic Tokens, agrupados por evento

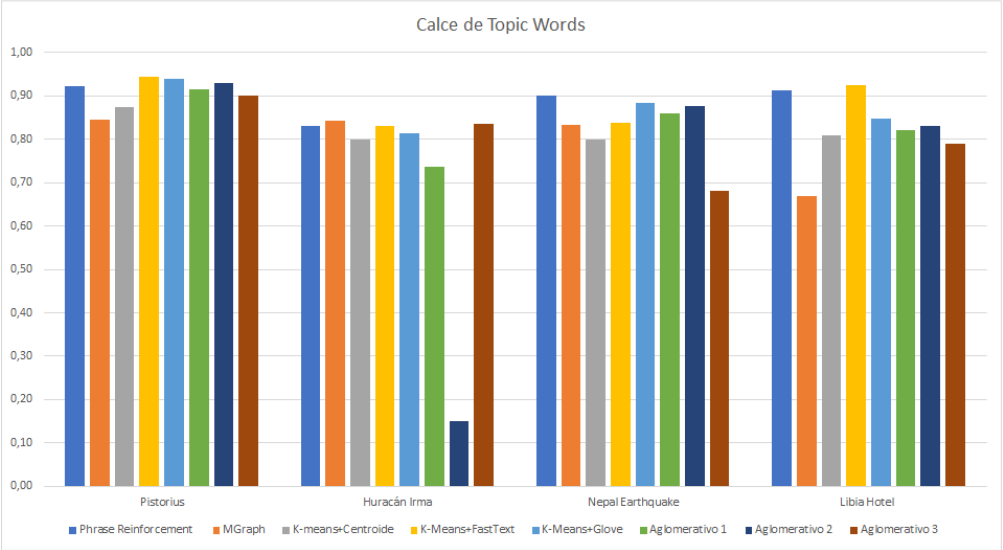

Figura 6.8: Resultados para Calce de Topic Words, agrupados por evento

Análisis

Esta evaluación si bien, permite establecer la calidad de un resumen en cuanto a cobertura, hay que tener algunas consideraciones al momento de analizar los resultados.

- Al evaluar la cobertura se favorecieron los resúmenes más largos, es decir, aquellos que tengan mayor cantidad de tweets o mensajes más largos. Esto porque al tener más texto disponible aumenta la probabilidad de que un término aparezca más veces, aumentando así la cobertura.

- En el caso de tweets que sean spam, estos también estaban en el conjunto inicial, por

lo cual su presencia en el resumen no afecto el valor de las métricas.

- El método de evaluación en su marco teórico no considera la existencia de spam. Sin embargo, en su aplicación real en publicaciones de redes sociales son implícitos al contenido.

Análisis de desempeño

Para los diferentes métodos se observaron algunas características como:

Phrase Reinforcement + K-means Este método fue el que obtuvo el mejor desempeño en todos los eventos. En efecto, para calce de topic words (figura 6.8) obtuvo un calce superior al 80 % en todos los eventos en comparación a los otros métodos que obtuvieron un calce promedio de 70 %, para en el caso de la divergencia de Jesen-Shannon (figura 6.5) obtuvo el menor valor generando una distancia del 10 % con quienes le siguen, recordemos que para está métrica entre menor sea el valor, mejor es el resultado.

La mejor cobertura que obtiene el método se debe a que permite generar resúmenes más largos que logra obtener mediante la generación de nodos que al enlazarce generan un camino más largo lo que produce un resumen mayor. El efecto de un resumen más grande es que aumenta la probabilidad que un término existente en conjunto original este presente el resumen.

MGraph Este método obtuvo distintos niveles de desempeño para distintos eventos. Por ejemplo para porcentaje de topic words, en la figura 6.7 obtuvo un 45 % de calce, en cambio para huracán irma obtuvo sobre un 80 % de calce. Esto se debe a que la calidad del resumen generado está directamente relacionada con la cantidad de imágenes que este evento contenga, ya que como se explicó anteriormente el resumen es generado solo sobre el conjunto de tweets con imágenes. Es por esto que este método obtiene mejores resultados en eventos más grandes como el terremoto en Nepal o el huracán Irma, en cambio con el evento de oscar pistorius no obtiene muy buenos resultados, ya que solo había 911 imágenes disponibles versus las 14.702 disponibles para el huracán Irma.

Word embeddings + Clustering Las distintas variaciones del método tuvieron desempeños distintos para los eventos. Algunas variantes tuvieron mejor desempeño. De las 6 variantes mostradas, las que obtuvieron mejores resultados fueron k-means+FastText y K-means+Glove, esto se aprecia en la figura 6.5 donde obtiene un desempeño similar a Phrase Reinforcement y en la figura 6.8 donde k-means+FastText obtiene el mejor desempeño. Por otra parte los que obtuvieron el peor desempeño fueron Aglomerativo 3 y Aglomerativo 1. En el caso de K-means con FastText, los mejores resultados se deben a que FastText aprende las representaciones vectoriales de los términos en base a sus n-gramas, esto tiene la ventaja que, al existir palabras infrecuentes o potencialmente mal escritas, estas aún pueden ser descompuestas en n-gramas los cuales pueden ser similares a los de términos más comunes, esto es muy útil para los mensajes de redes sociales donde abundan los términos mal escritos. Además, tanto para Glove como FastText consideran la información contextual presente al momento de generar la representación vectorial, es decir, si dos palabras distintas están rodeadas por los mismos términos, entonces la representación de dichas palabras sera similar.

Por otra parte los métodos basado en clustering aglomerativo obtuvieron un peor resultado, independiente de la representación vectorial utilizada, a partir de los gráficos de las figuras 6.6 y 6.5 vemos que Aglomerativo 3 obtuvo el peor desempeño, si . Esto se debe principalmente a que este tipo de clustering tiende a generar un clúster muy grande, que concentra la gran mayoría de los mensajes y el resto clúster mucho más pequeños, esto afecta la cobertura, porque existen tópicos que desaparecen o se confunden con otros. Esta distribución de los clusters se debe a la alta similitud que existe entre los mensajes de un evento. Esta distribución del tamaño de los clusters produce otro problema, si los clusters son muy pequeños el ranking por impacto propuesto en el método no funciona, ya que solo se tomará el tiempo de un par de mensajes y no será representativo.

6.1.3. Evaluación de diversidad

El objetivo de esta evaluación es cuantificar que tan similares son los mensajes que existen dentro de un mismo resumen.

Resultados

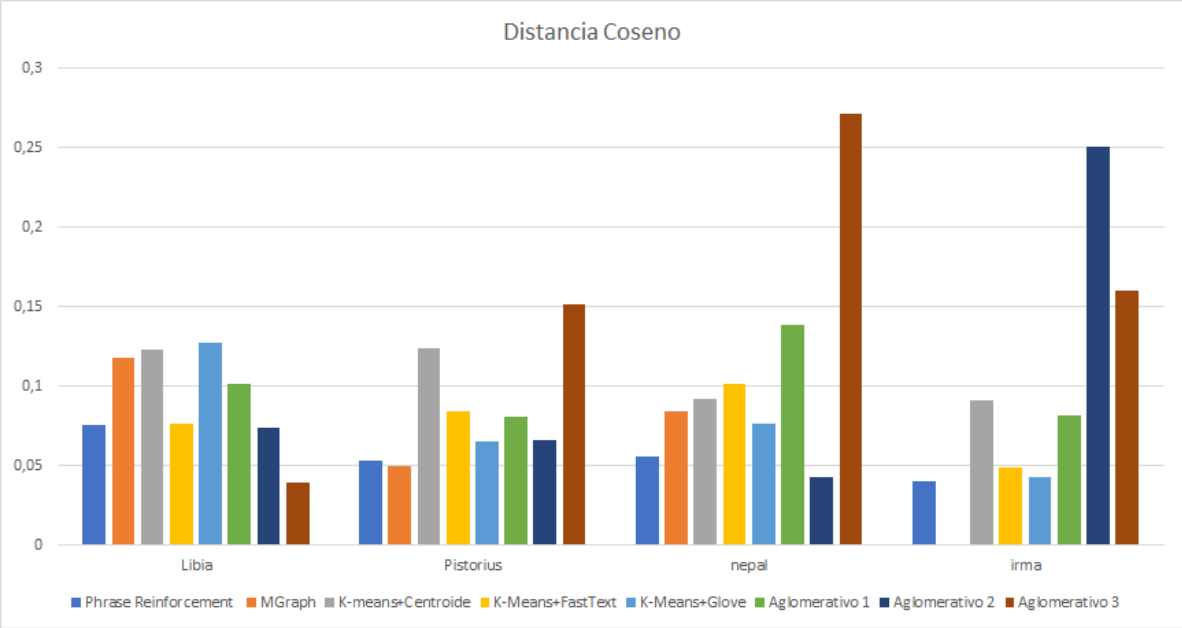

Figura 6.9: Resultados promedios para similitud coseno, agrupados por evento

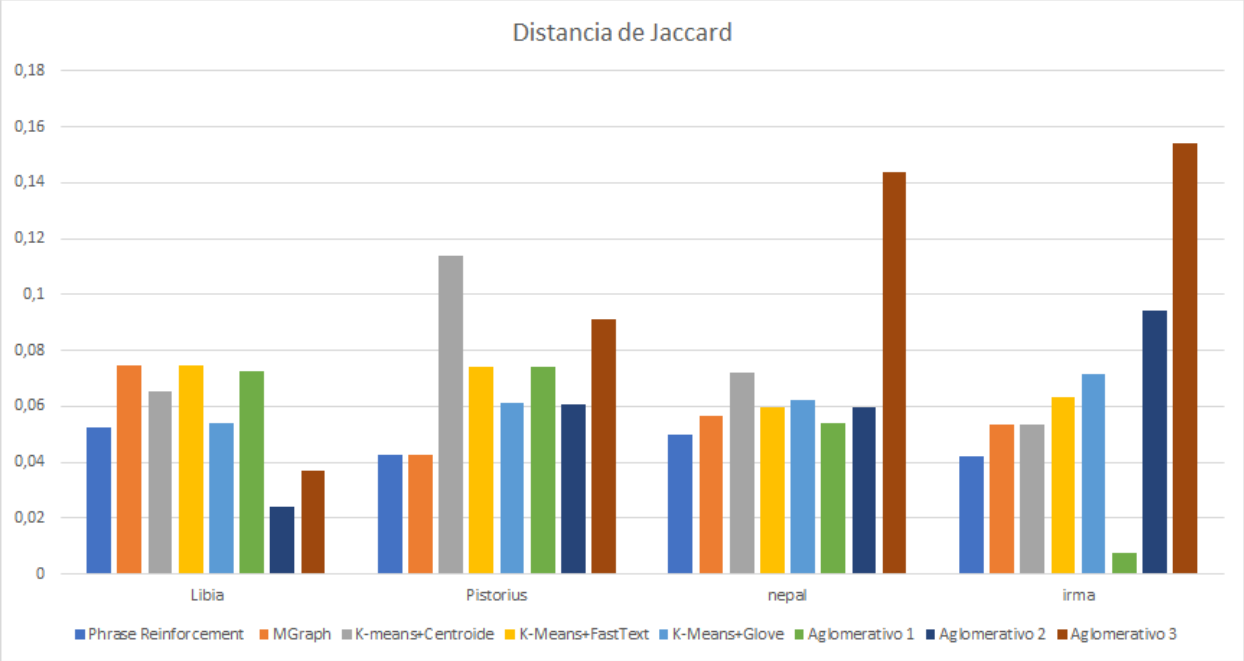

Figura 6.10: Resultados promedios para índice de jaccard, agrupados por evento

Análisis de resultados

Es importante notar, que puede darse la situación en que un resumen tenga muy buenos puntajes en esta evaluación, pero no contenga información importante, por ejemplo, el resumen puede contener solo mensajes con spam que sean todos distintos. Es por esto que esta evaluación debe ser acompañada por otra más que corrobore la calidad del resumen.

En general todos los resúmenes generados automáticamente presentan buenos puntajes en esta evaluación, más bajo es mejor, por lo cual se puede deducir que los mensajes dentro de un mismo resumen son distintos.

Para este caso el el método que obtuvo mejor desempeño fue Aglomerativo 1 obteniendo, por ejemplo un valor de indice de Jaccard (figura 6.10) menor al 10 %, con respecto a Phrase Reinforcement que obtuvo un indice de Jaccard del 12 % aproximadamente para todos los eventos. El desempeño de este método y de K-Means+FastText, se debe a la forma de agrupar los mensajes, al agrupar los mensajes por las URLs

El método que tuvo peor desempeño fue aglomerativo 3, obteniendo como se observa en la figura 6.9, hasta el doble con respecto a métodos como k-means+FastText para métricas como distancia coseno. Lo mismo sucede para el índice de Jaccard mostrado en la figura 6.10, en la cual este mismo método obtuvo un 40 % de calce comparado con Aglomerativo 1 que obtuvo cerca de un 5 % de calce. Este desempeño se justifica por los clusters generados. Al diferenciar la distribución de los mensajes por los distintos clusters, se observo que existían 9 de 15 clusters que contenían un solo mensaje. Esto indica que el método de clustering aglomerativo no es un buen método para agrupar los mensajes similares.

6.2. Análisis cualitativo de casos

Resultados

A continuación, se presentan los resultados obtenidos para cada evento y según los distintos métodos. Para esta evaluación solo se presentan los resultados de K-Means Fasttext (K-Means F), ya que todas las variantes generadas están fue la que obtuvo mejores resultados.

Tópicos	MGraph	Kmeans + PR	Kmeans	K-Means F
Carro bomba explota fuera del hotel	3	2	3	1
Isis proclama atentado	0	6	4	1
Número de Muertos en el atentado	6	8	7	9
Toma de Rehenes en el hotel	8	3	3	5
Número de atacantes	4	1	1	0
Enfrentamientos con fuerzas de seguridad libias	0	0	1	0
Spam	5	4	4	6
No relevantes	0	0	1	0

Tabla 6.2: Resultados de evaluación por tópicos para ataque en Libia

Tópicos	**MGraph**	**Kmeans + PR**	**Kmeans**	**K-Means F**
Oscar Pistorius pide disculpas a la familia de la victima	1	1	0	0
Oscar Pistorius vomita en la corte	2	1	1	1
Zombie Stopper	0	0	0	0
Policias revisan mensaje de whatsapp entre la pareja	0	0	0	0
Reeva Steenkamp le escribe una carta de san Valentín	0	0	0	0
Cricket bat vs gunshots	0	0	0	0
Oscar Pistorius se quita la prótesis durante el juicio	0	0	0	0
Oscar Pistorius es sometido a pruebas psiquiátricas	1	5	6	4
Argumentos finales del juicio	8	1	2	1
Pistorius alega inocencia	0	1	1	1
Casa de apuesta Paddy Powers realiza apuestas sobre el verdicto del juicio	0	1	1	0
Declaraciones de diversos testigos	4	3	2	3
Policía bajo investigación por manipular pruebas	0	0	0	0
Todas las sesiones de interrogatorio de Oscar Pistorius	1	2	3	6
Oscar pistorius es acusado de sacar y disparar una pistola en un restaurant	0	1	0	1
Spam	0	0	0	0
No informativo	10	7	8	3

Tabla 6.3: Resultados de evaluación por tópicos para el juicio de Oscar Pistorius

Tópicos	MGraph	Kmeans + PR	Kmeans	K-Means F
Avalancha Mt Everest	1	3	1	2
Conteo de Muertos por el sismo	11	12	8	5
Se reporta un sismo magnitud 7.8	6	5	3	7
Rescate de Personas entre los edificios	3	0	2	0
Formas de Ayudar a los damnificados	2	1	3	2
Ayuda Internacional entregada por distintos países	0	0	0	0
Daños a Edificios Históricos	2	1	1	1
Crisis Humanitaria	0	1	0	0
Destrucción de Edificios	3	3	1	2
Replicas que siguieron al sismo principal	0	1	0	3
No informativos	5	3	5	3
Spam	0	0	0	0

Tabla 6.4: Resultados de la evaluación por tópicos para el terremoto en nepal

Tópicos	**MGraph**	**Kmeans + PR**	**Kmeans**	**K-Means F**
Llegada a Florida del Huracán	6	4	4	5
Paso del huracán por paises del caribe	2	0	0	0
Evacuación del estado de Florida	0	0	0	1
Rescate de animales abandonados	0	0	1	1
Formación de 3 huracanes en el caribe	0	0	0	0
Reportes de inundaciones en distintas ciudades	2	1	0	0
Información oficial publicada por las autoridades	2	2	2	1
Recomendaciones sobre que hacer ante emergencias	3	0	1	1
Como ayudar a los damnificados por el huracan	1	0	0	0
Información sobre el huracán, como velocidad y dirección	7	4	4	2
Recogida del océano en las costas de florida	0	1	2	0
Criticas a las autoridades por manejo de la crisis	0	2	1	1
Reportes de cortes de luz	0	1	1	1
Reportes de saques en distintas ciudades	0	0	1	0
Información sobre los vuelos y estado de los aeropuertos	2	1	0	0
Spam	0	0	0	0
No relevantes	5	5	7	6

Tabla 6.5: Resultado de evaluación por tópicos para huracán Irma

Evento	MGraph	Kmeans + PR	Kmeans	K-Means F
Ataque en Libia	66,67 %	83,33 %	100,00 %	66,67 %
Juicio Oscar Pistorius	40,00 %	60,00 %	46,67 %	46,67 %
Terremoto en Nepal	70,00 %	80,00 %	70,00 %	70,00 %
Huracán Irma	53,33 %	53,33 %	60,00 %	53,33 %

Tabla 6.6: Porcentaje de Tópicos presentes para cada evento

Análisis de los Resultados

A partir de todas las tablas de la sección, podemos observar que existe una alta redundancia de tópicos para todos los resúmenes. Esto se contrapone a lo observado en la evaluación automática de diversidad, donde todas las métricas indican una alta diversidad. Esto es porque muchos mensajes parafrasean lo que ya fue dicho por otros mensajes, este parafraseo no alcanza a ser distinguido por ninguna representación vectorial, por lo cual tweets que hablan del mismo tema, quedan en clusters distintos. Un caso particular de esto es el conteo de muertos en el terremoto en Nepal, ya que aparte de utilizar distintos sinónimos para "death", el número de muertos estaba en constante actualización, esto explica los resultados vistos en 6.4, donde se observa, por ejemplo, que para el método phrase reinforcement más de la mitad de los mensajes mencionan el conteo de muertos.

Como se observa en la tabla 6.6, Phrase Reinforcement es el que obtiene los mejores resultados para todos los métodos, lo cual concuerda con lo observado en la evaluación automática de cobertura. Pero aun así presenta una alta redundancia de tópicos, por ejemplo, para el atentado en libia de 8 mensajes seleccionados mencionaban el tema "Número de muertos".

Si observamos la tabla 6.2, vemos que el evento de Libia presenta de forma transversal para todos los métodos, varios mensajes categorizados como spam, esto debido al problema que se explicó anteriormente con la recolección de los datos. Lo cual baja considerablemente la calidad del resumen. Además, al ser este un evento muy acotado en duración, ubicación y entidades relacionadas, podemos observar que la cantidad de tópicos existentes también es pequeña, por lo cual, al seleccionar una cantidad de mensajes mayor, se tenderán a repetir los tópicos, el problema es que al disminuir la cantidad de clusters y por consiguiente la cantidad de tópicos, se observa que se pierde más información, ya que se siguen formando clusters sólo con mensajes de spam.

El método K-Means DF tiene una alta redundancia en los tópicos, pero en la evaluación automática de cobertura y de diversidad obtiene muy buenos puntajes, sobre todo para el evento en libia, esto demuestra que para los eventos existe un gran número mensajes que hablan sobre lo mismo, pero de distinta forma, la ventaja que tuvo está variante fue que disminuyo la cantidad de mensajes irrelevantes seleccionados, sobre todo en eventos como pistorius con un 12 % y Nepal con un 12 % de mensajes irrelevantes, en comparación con los otros métodos.

Por otra parte, en la tabla 6.3, vemos que no hay mensajes de spam, pero si hay varios

mensajes catalogados como irrelevantes, esto es, porque existen entre un 12 % y 50 % de mensajes que buscan solo que los usuarios hagan click en la url que comparte, por ejemplo, el tweet mostrado en 5.3, menciona el evento, pero no aporta información. Lo mismo sucede con el terremoto en Nepal y el huracán Irma, al ser estos eventos asociados a desastres, producen un alto impacto en los usuarios, los cuales expresan sus sentimientos hacia el evento, más que compartir información del mismo.

Todos los métodos tuvieron un desempeño más bajo para este evento, existiendo muchos tópicos que, si bien estaban presentes en los mensajes, no fueron identificados por el clustering, esto se debe a que probablemente dichos mensajes eran demasiado similares a otros ya existentes, además hay que considerar que solo se está seleccionando un representante por tópico, pueden existir clúster que mezclen tópicos y no se aprecie en los representantes.

Después de este estudio, se puede responder parcialmente la pregunta número cuatro de investigación, que hace referencia a cuales son las técnicas y representaciones más efectivas para los resúmenes. A partir de los métodos utilizados y de los eventos escogidos se tiene que la mejor representación es Fasttext, ya que presento el desempeño más parejo para los distintos eventos. Para responder esta pregunta de forma integra se requiere considerar un mayor número de eventos, para determinar si esta tendencia se mantiene.

Capítulo 7

Conclusión

El objetivo de este estudio fue determinar cuales son las técnicas existentes para realizar resúmenes en redes sociales; cuales representaciones de datos resultan más efectivas y cuales son las métricas utilizadas para evaluar la calidad de un resumen.

Para determinar cuales son las técnicas existentes en la literatura, se realizo una revisión sobre los trabajos asociados a la generación de resúmenes. A partir de está se propone una nueva categorización de los trabajos enfocada particularmente en resúmenes automáticos sobre redes sociales. La categorización propuesta agrupa los trabajos en tres grupos distintos en base a las técnicas que utilizan para agrupar los mensajes y detectar los subeventos.

Esta revisión de la literatura permitió responder la pregunta de investigación número uno, planteada en el capitulo 1. Las métricas utilizadas para realizar las evaluaciones son muy similares, un gran número de los trabajos revisados utiliza ROUGE como métrica de evaluación. Aunque los trabajos utilizan esta métrica como base para determinar la calidad de un resumen, las evaluaciones realizadas entre los trabajos no son comparables. La dificultad para realizar dicha comparación radica en que ROUGE se basa en comparar el resumen generado con uno generado de forma manual, esto dificulta las comparaciones, porque los resúmenes producidos de forma manual casi siempre serán distintos, esto porque los individuos que generan los resúmenes son distintos y pueden generar el resumen considerando distintos aspectos. En conjunto con eso, los métodos revisados utilizan distintos conjuntos de datos para las evaluaciones, esto dificulta poder comparar dos métodos distintos.

Para responder la segunda pregunta de investigación presentada en el capitulo uno se realizo una comparación entre los vocabularios utilizados entre los mensajes publicados en redes sociales y lo publicado en medios tradicionales de noticias con respecto a un conjunto de 4 eventos distintos. A partir de esta comparación se determino que los vocabularios usados en ambos conjuntos son distintos, el porcentaje de intersección no supera el 10 %. Esto se debe principalmente a que lo que genera más discusión en las redes sociales no necesariamente es lo mismo que es presentado en los medios tradicionales. Esto se expresó en los distintos eventos, por ejemplo, al revisar las noticias relacionadas con el ataque terrorista en Libia dichos artículos, además de informar del evento mismo, explicaban el contexto político y social de Libia en ese instante, temas que no fueron discutidos en las redes sociales. Además,

a partir de esto se concluye que no se pueden usar este tipo de medios para evaluar resúmenes automáticos generados con mensajes de redes sociales.

Una vez implementados los métodos de las distintas categorías y generados los resúmenes de los respectivos eventos, se realizo un estudio comparativo entre los distintos métodos. A partir de este estudio se observo que existen dos factores que afectan la generación de los resúmenes:

- El primer factor, que influyen en la generación de un resumen, está presente en la detección y recolección de eventos, si la detección de eventos extrae mensajes que no estén relacionados con el evento, estos se pueden mezclar con los que si están asociados al evento y disminuir la calidad del resumen. Incluso si son muy numerosos pueden formar sus propio clúster de mensajes. Esto fue observado en el evento del ataque terrorista en Libia, donde una de las tuplas de búsqueda fue $(luxury, hotel)$, produciendo que se descargaran muchos mensajes asociados a hoteles de lujo y no sobre el ataque.

- El segundo factor, que afecta el resultado final del resumen, es la representación vectorial que se utilice para modelar el contenido textual de los mensajes. Por ejemplo, los métodos que utilizan fast-text y Glove como representación vectorial, obtuvieron mejores resultados. Esto porque, tanto fast-text como Glove son capaces de utilizar la información contextual que rodea los términos, es decir, dos palabras distintas podrían tener representaciones similares, porque los términos alrededor de dichas palabras son los mismos. En particular Fast-text obtiene mejores resultados, porque fast-text aprende las representaciones tanto de una palabra como de sus n-gramas, por lo cual si una palabra está mal escrita su representación vectorial sea similar a la de la palabra bien escrita.

Además para este estudio comparativo se utilizaron métricas aplicadas en la evaluación de generación de resúmenes de documentos tradicionales, comprobando que dichas métricas de diversidad y cobertura también son aplicables sobre conjuntos de resúmenes generados a partir de información de redes sociales, esto da un carácter más universal a dichas métricas.

Los factores anteriormente explicados, permiten responder la cuarta pregunta de investigación aunque solo de forma parcial, si bien los resultados obtenidos en este estudio comparativo son claros, la cantidad de eventos y métodos seleccionados no son suficientes para concluir que los dos factores explicados son relevantes para todo tipo de eventos, se debe realizar un estudio más amplio.

Durante el desarrollo de la investigación se determinó un espacio no cubierto por los métodos existentes, lo que da cabida a un nuevo método derivado de métodos existentes, que incluye dos consideraciones adicionales: La primera es agrupar los mensajes según las URLs que comparten y la segunda es poder armar un ranking de los subeventos detectados en base a su impacto. Este modelo se le denomino Word embbedings+Clustering, en reconocimiento a las técnicas usadas. Al aplicar las métricas de cobertura y diversidad, se comprobó que el método responde a la par respecto a los métodos ya establecidos.

Con lo expuesto anteriormente podemos afirmar que la hipótesis planteada en el capitulo 1. La información publicada en las redes sociales permite describir un evento noticioso de

forma completa. Ya que es posible generar resúmenes que cubran los aspectos más relevantes de cada evento utilizando esa información, esto se observa con el análisis de tópicos realizados, en donde se observa que para todos los eventos existe al menos un método que tiene al menos el 60 % de los tópicos presentes.

7.1. Trabajo Futuro

Una deficiencia que queda por cubrir para producir una mejora en los resúmenes es abordar con anterioridad la filtración de todos los mensajes que producen ruido, como Spam y mensajes sin contenido respecto al evento. El desafió está radicado en definir un mecanismo automático que permita generar un subconjunto sobre el cual trabajar que este ausente de estos ruidos.

Un punto importante a desarrollar seria realizar un estudio exhaustivo que permita responder la pregunta de investigación número cuatro. Para realizar este estudio se debería considerar un número mayor de eventos. También se debería realizar, en conjunto con la evaluación automática, un estudio con usuarios en el cual cada usuario asocie una evaluación al resumen y que en forma sistemática y automática que asocie un peso relativo a la calidad del resumen. Esto permitiría validar correctamente si lo obtenido con la evaluación automática es consistente con la percepción de los usuarios.

Otras actividades complementarias a desarrollar son:

- Determinar nuevas métricas complementarias a las presentadas que permitan medir con mayor exactitud el desempeño de un método, a modo de ejemplo, una métrica asociada a la importancia de los mensajes seleccionados.

- Determinar como la presencia de contenido multimedia afecta la comprensión del evento. Para estos efectos se propone nuevamente obtener evaluaciones desde los usuarios, en la cual junto con presentar el texto resumen, se propone mostrar un conjunto de imágenes que el usuario pueda seleccionar como representativas.

Finalmente, la importancia del trabajo ira en aumento en la medida que cada vez más usuarios requieran obtener información desde las redes sociales, cuyo universo está en un constante crecimiento. La necesidad de obtener información al menor tiempo y con el mejor desempeño hará necesario la evolución en la aplicación de los métodos que permitan obtener los resúmenes. El aporte de este trabajo está dado desde su marco teórico hasta la presentación de los resultados de las implementaciones, que podrán ser de utilidad para futuras investigaciones.

Bibliografía

Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, and Frederic Stahl. A survey of data mining techniques for social media analysis. *arXiv preprint arXiv:1312.4617*, 2013.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

Pinaki Bhaskar, Somnath Banerjee, and Sivaji Bandyopadhyay. A hybrid tweet contextualization system using ir and summarization. In *INEX*, page 164, 2012.

Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, pages 216–228, 2015.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, pages 993–1022, 2003.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Leon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in neural information processing systems*, pages 585–592, 1995.

Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. *ICWSM*, pages 66–73, 2011.

Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics, 1997.

Brian Donegan. *The Most Unforgettable Moments of Hurricane Irma*, 2017. Fecha Acceso: 2017-11-09. URL `https://weather.com/storms/hurricane/news/hurricane-irma-most-unforgettable-moments`.

Lois L Earl. Part-of-speech implications of affixes. *Mech. Translat. & Comp. Linguistics*, 9 (2):38–43, 1966.

Harold P Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*,

pages 264–285, 1969.

Pumza Fihlani. *Oscar Pistorius trial: 10 key moments*, 2015. Fecha Acceso: 2017-11-09. URL `https://www.bbc.com/news/world-africa-29018522`.

Ruifang He, Yang Liu, Guangchuan Yu, Jiliang Tang, Qinghua Hu, and Jianwu Dang. Twitter summarization with social-temporal context. *World Wide Web*, 20(2):267–290, 2017.

David Inouye and Jugal K Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 298–306. IEEE, 2011.

Rana Jawad. *Libya hotel attack: Five foreigners among nine killed*, 2015. Fecha Acceso: 2017-11-09. URL `https://www.bbc.com/news/world-africa-31001094`.

K Sparck Jones et al. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12, 1999.

Karen Sparck Jones and Julia R Galliers. *Evaluating natural language processing systems: An analysis and review*. Springer Science & Business Media, 1995.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Janani Kalyanam, Mauricio Quezada, Barbara Poblete, and Gert Lanckriet. Prediction and characterization of high-activity events in social media triggered by real-world news. *PloS one*, 11(12):e0166694, 2016a.

Janani Kalyanam, Mauricio Quezada, Barbara Poblete, and Gert Lanckriet. Prediction and characterization of high-activity events in social media triggered by real-world news. *PloS one*, 11(12):e0166694, 2016b.

Benjamin King. Step-wise clustering procedures. *Journal of the American Statistical Association*, pages 86–101, 1967.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.

Eric Levenson. *A day-by-day look at Hurricane Irma's path ahead*, 2017. Fecha Acceso: 2017-11-09. URL `https://edition.cnn.com/2017/09/08/us/forecast-hurricane-irma-timeline/index.html`.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, 2004.

Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501. Association for Computational Linguistics, 2000.

Elena Lloret and Manuel Palomar. Text summarisation in progress: A literature review. *Artif. Intell. Rev.*, pages 1–41, January 2012. ISSN 0269-2821.

Rui Long, Haofen Wang, Yuqiang Chen, Ou Jin, and Yong Yu. Towards effective event detection, tracking and summarization on microblog data. In *WAIM*, pages 652–663. Springer, 2011.

Annie Louis and Ani Nenkova. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, pages 267–300, 2013.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, pages 159–165, 1958.

Stuart Mackie, Richard McCreadie, Craig Macdonald, and Iadh Ounis. On choosing an effective automatic evaluation metric for microblog summarisation. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 115–124. ACM, 2014.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. Oakland, CA, USA., 1967.

Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236. ACM, 2011.

Philip J. McParlane, Andrew James McMinn, and Joemon M. Jose. "picture the scene...";: Visually summarising social media events. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14. ACM, 2014.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–387. ACM, 2012.

Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–655. Association for Computational Linguistics, 2012.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Ani Nenkova and Rebecca J Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152, 2004.

Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, pages 384–385, 2010.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Claire Phipps. Oscar pistorius trial: the full story, day by day, 2014. Fecha Acceso: 2017-11-09. URL `https://www.theguardian.com/world/2014/oct/21/oscar-pistorius-trial-full-story-reeva-steenkamp`.

Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

Martin F Porter. An algorithm for suffix stripping. *Program*, pages 130–137, 1980.

Dragomir R Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, and Elliott Drabek. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 375–382. Association for Computational Linguistics, 2003.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, pages 919–938, 2004.

Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. ISBN 0-201-12227-8.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM, 2009.

Manos Schinas, Symeon Papadopoulos, Yiannis Kompatsiaris, and Pericles A Mitkas. Mgraph: multimodal event summarization in social media using topic models and graph-based ranking. *International Journal of Multimedia Information Retrieval*, pages 51–69, 2016.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 685–688, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858099`.

Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 533–542. ACM, 2013.

Peter HA Sneath, Robert R Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, pages 11–21, 1972.

Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Ioannis Yiannis Kompatsiaris, Grigorios Tsoumakas, and Ioannis Vlahavas. A comprehensive study over vlad and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, pages 1713–1728, 2014.

Tadej Štajner, Bart Thomee, Ana-Maria Popescu, Marco Pennacchiotti, and Alejandro Jaimes. Automatic selection of social media responses to news. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 50–58. ACM, 2013.

Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, pages 525–526. Boston, 2000.

Josef Steinberger and Karel Ježek. Evaluation measures for text summarization. *Computing and Informatics*, pages 251–275, 2012.

Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2007.

Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, pages 1301–1315, 2015.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, pages 236–244, 1963.

Jiejun Xu and Tsai-Ching Lu. Seeing the big picture from microblogs: Harnessing social signals for visual event summarization. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 62–66. ACM, 2015.

Jiejun Xu, Samuel D Johnson, and Kang-Yu Ni. Cross-modal event summarization: A network of networks approach. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1653–1657. IEEE, 2016.

Shize Xu, Liang Kong, and Yan Zhang. A cross-media evolutionary timeline generation framework based on iterative recommendation. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 73–80. ACM, 2013.

Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas AJ Schweiger. Scan: a structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 824–833. ACM, 2007.

Duan Yajuan, Chen Zhimin, Wei Furu, Zhou Ming, and Heung-Yeung Shum. Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 763–780, 2012.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang.

Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 745–754. ACM, 2011.

Zhao Yanrong. *Nepal 2015 earthquake: timeline of events*, 2015. Fecha Acceso: 2017-11-09. URL `https://www.telegraph.co.uk/news/world/china-watch/society/nepal-earthquake-timeline/`.

Anexos

Anexo A

Lista de trabajo relacionado

En la siguiente tabla se listan todos los trabajos revisados, para la generación de la categorización propuesta en el capítulo 3. La tabla es presentada de forma horizontal para facilitar su visualización.

Tabla A.1: Listado de trabajos revisado para categorización

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
1	Event Summarization using Tweets	Chakrabarti, D., & Punera, K	2010	Twitter	Eventos deportivos	Cerrados	Probabilistico Hidden Markov Models	No	No especifica	Comparación manual con gold standard
2	Towards real-time summarization of scheduled events from twitter streams	Zubiaga, A., Spina, D., Amigó, E., & Gonzalo, J	2012	Twitter	Eventos deportivos	Cerrados	Probabilístico	No	Divergencia de Kullback-Leibler para selección de mensajes	Comparación con medios de noticias
3	Twitter topic summarization by ranking tweets using social influence and content quality	Yajuan, D., Zhimin, C., Furu, W., & Ming, Z	2012	Twitter	Terremotos	Cerrado	Probabilistic	No	Ranking de mensajes	Comparación contra gold standard propio
4	Automatic Twitter Topic Summarization	Wen, D., & Marshall, G.	2014	Twitter	Eventos seleccionados al azar	Abierto	Probabilistico Hidden Markov Models	No	Determina la probabilidad que un tweet pertenezca a un subevento, luego selecciona aquellos que mensajes que maximicen diversidad y cobertura	Evaluación extrinsica con jueces humanos

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
5	Summarizing Microblogs Automatically	Beaux Sharifi, Mark-Anthony Hutton, & Jugal Kalita	2010	Twitter	Eventos seleccionados al azar	Abierto	Grafos	No	Ranking de mensajes según calce con el grafo	Comparación con gold standard propio
6	Twitterstand: news in tweets	Sankaranarayanan, J., Lieberman, M. D., Teitler	2009	Twitter	Noticias	Abierto	Online clustering	No	Clasificación de los mensajes para determinar su relevancia, agrupa los mensajes relevantes según ubicación geográfica, muestra los n tweets más recientes	No especifica
7	Comparing twitter summarization algorithms for multiple post summaries	D Inouye, JK Kalita	2011	Twitter	Eventos seleccionados al azar	Abierto	Clustering	No	Selección utilizando TF-IDF, considerando cobertura	Comparación con gold standard propio
8	Sumblr: Continuous Summarization of Evolving Tweet Streams	Lidan Shou& Zhenhua Wang	2013	Twitter	Eventos seleccionados al azar	Abierto	Clustering	No	Realiza clustering online sobre los mensajes, selecciona mensajes que maximicen la cobertura y diversidad	Comparación contra gold standard propio

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
9	Towards Effective Event Detection, Tracking and Summarization on Microblog Data	R Long, H Wang, Y Chen, O Jin, Y Yu	2012	Sina Weibo	Eventos seleccionados al azar	Abierto	Grafos	No	Selección de mensajes que maximicen la cobertura de cada subevento	Comparación con el conjunto inicial de datos considerando cobertura
10	TweetMotif: Exploratory Search and Topic Summarization for Twitter.	B O'Connor, M Krieger, D Ahn	2010	Twitter	No, basado en querys ingresada por el usuario	Abierto	Probabilístico	Si	Selección de tópicos según qué tan probable sea una frase en un conjunto de tweets	No especifica
11	Social context summarization	Z Yang, K Cai, J Tang, L Zhang, Z Su, J Li	2011	Twitter	No, genera resúmenes de documentos web a partir de tweets	Abierto	Grafos	No	Determina la importación de un tweet según distintas características	Comparación con gold standard propio
12	Multimedia summarization for social events in microblog stream	Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua	2015	Sina Weibo	Eventos seleccionados al azar	Abierto	Probabilístico Hidden Markov Models	No	Selección de mensajes e imágenes según cobertura y diversidad. Los criterios son definidos por los autores	Comparación con gold standard propio

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
13	Structured event retrieval over microblog archives	Donald Metzler, Congxing Cai, Eduard Hovy	2012	Twitter	Eventos seleccionados al azar	Abierto	Probabilístico	No	Ranking de mensajes considerando cobertura y peaks de actividad en un periodo de tiempo específico	Comparación con gold standard propio
14	Automatic Summarization of Events From Social Media	Freddy Chong Tat Chua	2013	Twitter	Eventos predefinidos	Abierto	Probabilístico	No	Ranking de mensajes según cobertura	Comparación con páginas de wikipedia y gold standard propio
15	Experiments in microblog summarization	Beaux Sharifi, Mark-Anthony Hutton & Jugal K. Kalita	2010	Twitter	Trending topics seleccionados al azar	Abierto	Grafos	No	Selección de mensajes que contengan los términos más frecuentes dentro del conjunto de datos inicial	Comparación contra gold standard propio
16	MGraph: multimodal event summarization using topic models and graph-based ranking	Schinas, M., Papadopoulos, S., Kompatsiaris, Y., & Mitkas, P. A	2016	Twitter	Eventos seleccionados al azar	Abierto	Grafos	Si	Ranking de tweets con imágenes considerando cobertura, diversidad e impacto social del mensaje	Evaluación de calidad de texto con jueces humanos

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
17	Adaptive Representations for Tracking Breaking News on Twitter	Brigadir, I., Greene, D., & Cunningham	2014	Twitter	Timelines seleccionadas al azar	Abiertos	Probabilístico	No	Detección de eventos y generación automática de timelines	Evaluación con timelines generados manualmente, considerando cobertura y diversidad
18	Cross-modal event summarization: A network of networks approach	Jiejun Xu, Samuel D. Johnson	2016	Twitter	Solo usa un evento, considera noticias e imágenes	Especifico	Grafos	No	Selecciona mensajes, noticias e imágenes considerando la similitud e importancia de cada elemento	Evaluación manual de los timelines generados considerando cobertura y divdersidad
19	Automatic Selection of Social Media Responses to News	Tadej Stajner, Bart Thomee & Ana-Maria Popescu	2013	Twitter	Eventos noticiosos seleccionados al azar	Abierto	Otro (optimización)	No	Selecciona mensajes que minimicen la entropía, considerando aspectos textuales, sociales y características de los autores	Comparación contra gold standard propio
20	Summarizing a Document Stream	Hiroya Takamura, Hikaru Yokono& Manabu Okumura	2011	Twitter	Eventos deportivos	Cerrado	Otro (optimización)	No	Seleccionar los mensajes que minimicen la redundancia considerando la fecha de publicación y limitando el largo de los mensajes	Comparación contra gold standard propio

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
21	Personalized time-aware tweets summarization	Zhaochun Ren, Edgar Meij, Shangsong Liang	2013	Twitter	Eventos seleccionados al azar	Abierto	Probabilístico	No	Selecciona mensajes de acuerdo con la similitud entre la distribución de probabilidades de los mensajes y la distribución del subevento, además considera una ventana máxima de tiempo por evento	Comparación contra gold standard propio, construido a partir de las interacciones de los usuarios
22	On Summarization and Timeline Generation for Evolutionary Tweet Streams	Zhenhua Wang, Lidan Shou, Ke Chen	2015	Twitter	Eventos seleccionados al azar	Abierto	Clustering	No	Realiza clustering online sobre los mensajes, selecciona mensajes que maximicen la cobertura y diversidad	Comparación contra gold standard propio
23	TSum4act: A Framework for Retrieving and Summarizing Actionable Tweets during a Disaster for Reaction	Minh-Tien NGUYEN & Asanobu Kitamoto	2015	Twitter	Desastres naturales	Cerrado	Grafos	No	Realiza un ranking de subeventos, considerando su similitud e importancia con respecto a otros subeventos	Mide cuantos mensajes de un conjunto de datos etiquetados manualmente son seleccionados por el algoritmo

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
24	Evaluating methods for summarizing Twitter posts	Gary Beverungen y Jugal Kalita	2011	Twitter	Eventos seleccionados al azar a partir de hashtags	Abierto	Clustering	No	Generación de clusters con los tópicos de un evento	Evaluación automática con respecto a la calidad de los clusters obtenidos, midiendo dispersión y distancia de los elementos de clúster
25	Experiments in microblog summarization	Beaux Sharifi, Mark-Anthony Hutton & Jugal K. Kalita	2010	Twitter	Trending topics seleccionados al azar	Abierto	Grafos	No	Selección de mensajes que contengan los términos más frecuentes dentro del conjunto de datos inicial	Comparación contra gold standard propio
26	Entity-centric topic-oriented opinion summarization in twitter	Xinfan Meng, Furu Wei, Xiaohua Liu & Houfeng Wang	2012	Twitter	Opiniones sobre marcas y personas	Cerrado	Otros	No	Clasifica los mensajes en 7 categorías relevantes, luego selecciona aquellos que maximicen relevancia y legibilidad	Comparación contra gold standard propio, obtenido a partir de los subeventos detectados en la parte anterior. Utiliza un caso de estudio

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
27	Generating event storylines from microblogs	Chen Lin, Chun Lin, Yang Chen, Tao Li & Jingxuan Li	2012	Twitter	Dataset provisto por TREC	Abierto	Probabilístico y grafos	No	Generación de timelines considerando cobertura con respecto a los tópicos relacionados	Comparación contra gold standard de TREC 2011
28	Why is sxsw trending?: exploring multiple text sources for twitter topic summarization	Fei Liu, Yang Liu & Fuliang Weng	2011	Twitter	Trending topics seleccionados al azar	Abierto	Otro (optimización)	No	Selecciona mensajes que maximicen función objetivo, considerando cobertura y relevancia	Evaluación de calidad de texto realizada de forma manual
29	Relevance Modeling for Microblog Summarization	Sanda Harabagiu & Andrew Hickl	2011	Twitter	Eventos seleccionados al azar	Abierto	Probabilístico	No	Determina la probabilidad que un tweet sea relevante o no para un subevento, luego selecciona los k tweets más recientes que tengan mayor probabilidad de ser relevantes	Evaluación de calidad de texto realizada de forma manual

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
30	Automatic twitter topic summarization with speech acts	Renxian Zhang, Wenjie Li, Dehong Gao & You Ouyang	2013	Twitter	Eventos escogidos al azar	Abierto	Otro	No	Genera resúmenes extractivos	Comparación con gold standard generado a partir de tweets validados por usuarios y evaluación manual de calidad de texto
31	Multimedia social event detection in microblog	Yue Gao, Sicheng Zhao, Yang Yang & Tat-Seng Chua	2015	Sina Weibo	Eventos escogidos al azar	Abierto	Grafos	No	Agrupación de mensajes segun cobertura, tiempo, impacto social y contenido visual, se seleccionan aquellos tweets con mayor grado de conexión	Comparación contra gold standard propio
32	Efficient Online Summarization of Microblogging Streams	Adrei Olariu	2014	Twitter	Eventos escogidos al azar	Abierto	Grafos	No	Generación de resúmenes abstractivos a partir de un grafo de palabras	Evaluación de calidad de texto con jueces humanos

Continua en la página siguiente

ID	Titulo	Autores	Año	Fuente de Datos	Usa Eventos	Tipos de datos	Técnica Principal	Código Fuente	Criterios de selección	Criterios de evaluación
33	Towards Effective Event Detection, Tracking and Summarization on Microblog Data	R Long, H Wang, Y Chen, O Jin, Y Yu	2012	Sina Weibo	Eventos seleccionados al azar	Abierto	Grafos	No	Selección de mensajes que maximicen la cobertura de cada subevento	Comparación con el conjunto inicial de datos considerando cobertura
34	Towards Twitter context summarization with user influence models	Y Chang, X Wang, Q Mei, Y Liu	2013	Twitter	Eventos seleccionados al azar	Abierto	Grafos	No	Selección de mensajes considerando la popularidad de un usuario, la diversidad de los mensajes e impacto del mensaje	Comparación con gold standard armado manualmente por los autores
35	Multi-tweet summarization of real-time events	Muhammad Khan, Danushka Bollegala y Guangwen Liu	2013	Twitter	Eventos seleccionados a partir de hashtags en intervalos de tiempo especifico	Específico	Grafos	No	Selección de mensajes utilizando pageRank para cada tópico detectado, además considera la diversidad entre mensajes	Comparación de los resumen generados con los obtenidos por dos baselines distintos