

# Detectando la prioridad de contenidos generados en Twitter por medio de $n$ -gramas de palabras

Gerardo García-Rodríguez, Gabriela Ramírez-de-la-Rosa,  
Esau Villatoro-Tello, Héctor Jiménez-Salazar

Departamento de Tecnologías de la Información,  
División de Ciencias de la Comunicación y Diseño,  
Universidad Autónoma Metropolitana, Unidad Cuajimalpa, DF, México

210368185@alumnos.cua.uam.mx,  
{gramirez,evillatoro,hjimenez}@correo.cua.uam.mx

**Resumen.** En la actualidad, el uso de las redes sociales ha revolucionado la forma en que los usuarios intercambian ideas, opiniones e información. Gracias a este cambio de paradigma en la forma de interactuar entre los usuarios, las grandes compañías y personajes públicos han comenzado a prestar particular interés a la opinión generada respecto a sus productos y/o servicios, actos y/o eventos dentro de las redes sociales. A esta actividad se le conoce como el Análisis de Reputación en Línea (ARL), actividad desempeñada principalmente por usuarios expertos en el análisis de imagen y que al mismo tiempo son capaces de tomar decisiones estratégicas que ayuden a mejorar la reputación de una empresa o personaje público. La relevancia que ha tomado esta actividad en últimos años ha motivado a la comunidad científica a proponer métodos automáticos que apoyen las labores de un ARL. En este trabajo se propone un método automático para determinar cuándo un tuit es importante dentro de una categoría predefinida de mensajes. El método propuesto se basa en el uso de  $n$ -gramas para establecer la importancia de los contenidos generados en Twitter. Los experimentos realizados muestran que la ocurrencia de ciertos términos permite a un modelo de clasificación automática determinar de forma efectiva (*medida-F=0.7*) cuando un tuit es o no importante para un ARL.

**Palabras clave:** Bolsa de  $n$ -gramas, clasificación de contenidos, clasificación no temática, procesamiento de lenguaje natural, aprendizaje automático.

## 1. Introducción

Hoy en día, el amplio uso de las redes sociales ha generado una nueva forma de intercambiar ideas, opiniones e información en tiempo real y desde cualquier parte del planeta, incluso desde fuera de él<sup>1</sup>. Estos intercambios de ideas e

<sup>1</sup> El Ingeniero astronauta T.J. Creamer envió, a través de Twitter, un mensaje desde la Estación Espacial Internacional en el 2010: <http://techland.time.com/2010/01/22/a-message-from-outer-space/>

información se llevan a cabo no solo entre usuarios comunes, sino que también involucra a usuarios que representan grandes empresas, artistas, figuras públicas, personajes políticos, por mencionar algunos.

A través del uso de Twitter, por ejemplo, un consumidor puede contactar de forma casi instantánea a la empresa de algún producto para reportar un fallo o para preguntar por un componente que necesita. De igual manera, la compañía puede saber de forma inmediata cómo está respondiendo un determinado grupo de usuarios ante la presencia de alguno de sus productos en el mercado; o si se piensa en algún personaje público, este puede conocer cómo sus seguidores están reaccionando ante una determinada campaña, etc.

Si bien las ventajas que las redes sociales ofrecen a las empresas son grandes, la cantidad de información producida en éstas es gigantesca. Consecuentemente, un analista de reputación enfrenta el gran problema de revisar y analizar toda la información producida en las redes sociales referentes a su producto/servicio o personaje público en cuestión, lo cual resulta en una actividad extremadamente abrumadora y que es realizada con poca eficiencia. Es importante mencionar que tan solo en Twitter se producen cerca de 500 millones de mensajes (tuits) al día [13]. El manejo manual de tal cantidad de información es prácticamente imposible. En este contexto, el proponer sistemas automáticos que permitan realizar un Análisis de Reputación en Línea (ARL) es necesario para procesar y entender las grandes cantidades de contenidos generados en Twitter asociados a una compañía o una entidad en particular [1].

Este análisis de reputación en línea se basa en el constante monitoreo de los mensajes producidos en las redes sociales. El monitoreo de reputación en línea se realiza en etapas. La primera etapa del proceso es filtrar sólo los mensajes que pertenecen a la entidad (compañía o persona); posteriormente, el analista debe decidir la polaridad de dichos mensajes, *i.e.* las implicaciones positivas o negativas que dichos mensajes podrían producir a la compañía o persona. Debido a que una compañía no solo maneja un único producto, así como los personajes públicos se desenvuelven en diferentes ámbitos, el siguiente paso del proceso de ARL consiste entonces en agrupar los mensajes de acuerdo a temáticas relacionadas a un determinado producto o servicios. Finalmente, una vez que se han identificado las temáticas abordadas en los mensajes, el analista deberá determinar cuáles de esos mensajes son los realmente importantes para la compañía o en cuáles debe prestarse mayor atención debido a las implicaciones que podría generar al interior de la compañía, o en el caso de una persona, implicaciones en la imagen pública de dicho personaje. Note que para determinar la importancia de estos mensajes es necesario conocer el giro de la empresa o ámbito en el que el personaje se desenvuelve, pues de ésto dependerá la forma en que el analista emita un juicio adecuado.

Cada uno de los procesos mencionados anteriormente tienen sus propios desafíos, *e.g.*, desambiguación de palabras, análisis de opinión, análisis de sentimientos, agrupamiento, etc. Dada la complejidad que implica realizar todas las tareas involucradas en el ARL, en este trabajo nos enfocaremos únicamente en la detección automática de la importancia de cada mensaje; es decir, una

vez que los tuits están agrupados por temáticas, nos interesa detectar de forma automática cuáles son los tuits de mayor prioridad que un analista deberá tener presente para su análisis. A este último proceso del monitoreo de reputación en línea también se le llama *detección de prioridad* [1].

El método propuesto en este artículo se basa en un enfoque de clasificación de textos tradicional. Nuestra hipótesis principal establece que la importancia de un tuit puede ser determinada a partir de las palabras contenidas en él y que de alguna forma reflejan la forma y uso de ciertos términos que son dependientes del giro o categoría de la compañía y/o personaje que se este analizando. Los resultados obtenidos en 4 diferentes categorías muestran que un enfoque tradicional de clasificación de textos aporta elementos suficientes para determinar cuándo un tuit es importante dentro de una determinada categoría.

El resto del artículo está organizado como sigue, en la Sección 2 se describirán algunos trabajos evaluados dentro de la competencia RepLab [9] que abordan el problema de detección de prioridad. El método propuesto se describirá en detalle en la Sección 3. En la Sección 4 se describe el marco experimental usado para evaluar el método propuesto y finalmente se presentarán conclusiones y trabajo futuro en la Sección 5.

## **2. Trabajo relacionado**

Actualmente, sistemas automáticos para el monitoreo de reputación en línea pueden ser evaluados en el contexto de la competencia internacional RepLab [9]. En dicho espacio se presentan métodos que abordan cada uno de las etapas involucradas en monitoreo de reputación en línea descritas en la sección anterior. Debido a que en este artículo se propone un sistema automático para detectar la prioridad de los tuits, a continuación se analizarán de forma breve los sistemas dentro de la competencia RepLab que se enfocaron en este problema.

La mayoría de los sistemas que atacan el problema de detección de prioridad en tuits utilizan técnicas de análisis de sentimientos, bajo el supuesto de que tuits de connotación negativa hacia la compañía deberán ser más importantes que tuits de connotaciones positivas [3,7,12]. Por ejemplo, en [12] hacen uso de un diccionario de términos afectivos llamado *SentiSense*, el método que proponen genera un grafo para cada dominio con el objetivo de expandir este diccionario de palabras de afecto.

En [3] asignan la prioridad de cada tuit basándose en la polaridad y la intensidad de sentimiento en ellos, debido a que no contaban con un diccionario de sentimiento especial para redes sociales, en particular Twitter, utilizaron un diccionario inicialmente creado para noticias. Sin embargo, el uso de sentimiento (de noticias) no ayudó a la identificación de la prioridad de los tuits.

A diferencia de los trabajos anteriores, el método propuesto en este trabajo no busca utilizar técnicas de análisis de sentimiento para encontrar la prioridad de los tuits, ni hace uso de recursos externos para apoyar la clasificación. Por el contrario, nuestro método entrena un modelo para asignar prioridades por categoría utilizando un enfoque tradicional de clasificación de textos, como se

explicará en la Sección 3. Los experimentos realizados en cuatro diferentes categorías muestran que un enfoque de clasificación de textos tradicional aporta suficientes elementos para seleccionar tuits importantes.

### 3. Método propuesto

Como se ha venido mencionando en secciones anteriores, en este trabajo se aborda la problemática de detección de tuits importantes aplicando el paradigma de Clasificación de Textos<sup>2</sup>. Bajo este paradigma un primer paso necesario es el *indexado* (*i.e.*, representación) de los documentos de entrenamiento ( $Tr$ ), actividad que denota hacer el mapeo de un documento  $d_j$  en una forma compacta de su contenido. La representación más comúnmente utilizada para representar cada documento es un vector con términos<sup>3</sup> ponderados como entradas, concepto tomado del modelo de espacio vectorial usado en recuperación de información [2]. Es decir, un texto  $d_j$  es representado como el vector  $\vec{d}_j = \langle w_{kj}, \dots, w_{|\tau|j} \rangle$ , donde  $\tau$  es el *diccionario*, *i.e.*, el conjunto de términos que ocurren al menos una vez en algún documento de  $Tr$ , mientras que  $w_{kj}$  representa la importancia del término  $t_k$  dentro del contenido del documento  $d_j$ .

En ocasiones  $\tau$  es el resultado de filtrar los términos del vocabulario, este filtrado normalmente es el resultado de un preprocesamiento (ver Sección 3.2 para más detalles). Una vez que hemos hecho los filtrados necesarios, el diccionario  $\tau$  puede definirse de acuerdo a diferentes criterios, sin embargo el que se empleó en esta propuesta corresponde a la Bolsa de Palabras (BOW<sup>4</sup>).

La BOW es la forma tradicionalmente utilizada para representar los documentos [11]. Este método de representación utiliza a las palabras simples como los elementos del vector de términos. Dentro de este trabajo se evaluó la influencia tanto de palabras simples (unigramas de palabras) así como de secuencias de palabras de longitud dos (bigramas de palabras).

Respecto al peso  $w_{kj}$ , existen diferentes formas de calcularlo, entre las más usadas en la comunidad científica se pueden mencionar: el ponderado booleano, ponderado por frecuencia y ponderado por frecuencia relativa de términos. A continuación se describen de manera breve cada uno de estos esquemas de pesado.

#### 3.1. Esquemas de pesado utilizados

El peso  $w_{kj}$  puede ser calculado utilizando diferentes enfoques, el más simple de estos enfoques es el *booleano* que consiste en asignar un valor de 1 al término si éste aparece en el documento, y 0 en caso contrario. Este esquema de pesado

<sup>2</sup> La Clasificación de Textos es la tarea de asociar automáticamente categorías predefinidas con documentos a partir del análisis de su contenido [11].

<sup>3</sup> Entiéndase por términos ya sea palabras,  $n$ -gramas de palabras y/o caracteres, secuencias frecuentes, etc.

<sup>4</sup> Bag Of Words por sus siglas en Inglés.

captura de forma intuitiva la presencia y/o ausencia de los términos dentro de un documento.

Otro esquema de pesado es conocido como *frecuencia relativa*. En este enfoque se determina el peso de un término  $t_k$  en el documento  $d_j$  en proporción directa al número de veces que el término aparece en el documento, e inversamente proporcional al número de documentos en los que aparece el término  $t_k$  el conjunto total de entrenamiento. En particular, el peso está dado por:

$$w_{kj} = TF(t_k) \times IDF(t_k) \quad (1)$$

donde  $TF(t_k) = f_{kj}$ , es decir, la frecuencia del término  $t_k$  en el documento  $d_j$ . Y la *frecuencia inversa*, o  $IDF(t_k)$  es una manera de medir la *rareza* del término  $t_k$ . Para calcular el valor de IDF se utiliza la siguiente fórmula [10]:

$$IDF(t_k) = \log \frac{|Tr|}{\{d_j \in D : t_k \in d_j\}} \quad (2)$$

A esta frecuencia relativa le denominaremos *Tf-Idf* a lo largo de este documento.

Ambos esquemas han demostrado tener buenos resultados en métodos de clasificación de textos. El primer esquema es simple y sólo considera la aparición o no de términos en el documento, mientras que el segundo método es más complejo de calcular pero ayuda a discriminar términos que no apoyan la clasificación de una u otra clase. *Tf-Idf* prefiere términos que sean frecuentes en el documento pero que aparezcan poco en documentos del resto de las clases. En los experimentos evaluaremos estos dos esquemas por separado, nuestra intuición es que debido a lo corto de los mensajes de los tuits, el primer enfoque será suficiente.

### 3.2. Preprocesamiento

Un tuit es un texto de no más de 140 caracteres que puede contener componentes inherentes a esa red social como: palabras fuera de algún diccionario, texto informal, *hashtags*, menciones a otros usuarios, URLs, abreviaciones de palabras, repeticiones de letras, repeticiones de signos de puntuación, etc. La mayoría de estos componentes no son útiles en un enfoque de clasificación de textos tradicional [11,8] ni en la definición de la prioridad del mensaje [3,7,12]. Por lo anterior, cada uno de los mensajes han sido preprocesados, aplicando los siguientes pasos:

- Los tuits se convierten a minúsculas con la finalidad de normalizar el vocabulario.
- Cualquier secuencia de espacios en blanco se convierte en un solo espacio.
- Cualquier mención a usuarios encontrada es eliminada.
- Cualquier enlace (URL) encontrado es eliminado.
- Se eliminan los signos de puntuación. En consecuencia cualquier emoticono queda de igual forma eliminado.
- Se eliminan las palabras vacías y/o funcionales.

## 4. Evaluación experimental

A continuación se describen las medidas usadas en la evaluación del sistema, el conjunto de datos usados y los resultados obtenidos en los experimentos realizados.

### 4.1. Medidas de Evaluación

Para evaluar el método propuesto se utilizaron las medidas tradicionales para evaluación de sistemas de clasificación de textos, tales como *precisión*, *recuerdo* y *medida F*. La precisión ( $P$ ) es la proporción de textos clasificados correctamente en una clase  $c_i$  con respecto a la cantidad de textos clasificados en esa misma clase. El recuerdo ( $R$ ), la proporción de textos clasificados correctamente en una clase  $c_i$  con respecto a la cantidad de textos que realmente pertenecen a esa clase. Así, la precisión se puede ver como una medida de la corrección del sistema, mientras que el recuerdo da una medida de cobertura o completitud.

Normalmente se emplea la medida  $F$  para describir el comportamiento de la clasificación, la cual se define como:

$$F = \frac{(1 + \beta^2)P * R}{\beta^2(P + R)} \quad (3)$$

donde  $\beta$  representa la media armónica entre la precisión y el recuerdo. La función de  $\beta$  es la de controlar la importancia relativa entre las medidas de precisión y recuerdo. Es común asignar un valor de 1 indicando igual importancia a ambas medidas.

### 4.2. Conjunto de datos

En los experimentos presentados se usó la colección de tuits proporcionada por el RepLab 2013 [1,9]. La colección contiene tuits de cuatro grandes categorías, cada categoría busca cubrir un tipo diferente de evaluación de reputación. Las cuatro categorías son: *automotriz* que incluye entidades para las cuales la reputación se determina en base a sus productos; *banca* con entidades para las cuales la transparencia y la ética de sus actividades son factores a considerar en cuanto a la evaluación de su reputación; *universidades* donde su reputación depende de la variedad de productos intangibles; y *musical* en la que sus entidades basan su reputación tanto en la calidad de sus productos como en las cualidades personales de forma igualitaria.

Agregado a lo anterior, los tuits proporcionados por el RepLab 2013 están en dos idiomas distintos: Inglés y Español. Para la realización de nuestros experimentos se trabajó únicamente con el subconjunto de tuits en español (vea la Tabla 1).

En la Tabla 1 se muestra información sobre el subconjunto de datos usado en nuestros experimentos. Cabe mencionar que la colección de datos original menciona que los tuits están clasificados en tres categorías, a saber, *importante* ( $I$ ),

*medianamente importante (MI) y no importante (UN)*; sin embargo, es necesario mencionar que en los datos de entrenamiento liberados por los organizadores del RepLab 2013, no existe ningún ejemplo de tuit *importante* en ninguna de las cuatro categorías. Dada esta situación, nuestro problema se redujo a un problema de clasificación binaria, es decir, identificación de tuits *medianamente importantes y no importantes*.

**Tabla 1.** Número de tuits MI y UN existentes en cada una de las diferentes categorías en el subconjunto de datos considerado

|              | Automotriz | Banca | Musical | Universidades | Total |
|--------------|------------|-------|---------|---------------|-------|
| <i>MI</i>    | 952        | 540   | 1497    | 141           | 3130  |
| <i>UN</i>    | 1256       | 1011  | 1462    | 79            | 3808  |
| <i>Total</i> | 2208       | 1551  | 2959    | 220           | 6938  |

Note que de forma global el conjunto de tuits MI y UN está balanceado. Sin embargo, al ver el número de tuits por clase en cada categoría, especialmente para las categorías Automotriz, Banca y Universidades, no dice que estas clases se encuentran muy desbalanceados, situación que representa ya un reto por sí misma.

### 4.3. Clasificadores

Dado que nuestra propuesta para identificar la importancia de los contenidos generados en Twitter no depende de ningún algoritmo de aprendizaje en particular, podemos emplear prácticamente cualquier clasificador para enfrentar el problema. Para los experimentos realizados seleccionamos dos diferentes algoritmos de aprendizaje, los cuales son algoritmos representativos dentro de la gran variedad de algoritmos de aprendizaje disponibles actualmente en el campo de aprendizaje computacional [4,6]. Específicamente, consideramos los siguientes:

- **Naïve Bayes(NB).** Método probabilístico que asume la independencia de los atributos entre las diferentes clases del conjunto de entrenamiento.
- **Máquinas de Vectores de Soporte (Support vector machine, SVM).** Método que hace uso de discriminantes lineales que intentan localizar de forma óptima el hiperplano que separa los elementos de las diferentes clases. Para nuestro trabajo se utilizó un kernel lineal.

En nuestros experimentos se empleó la implementación de Weka [5] de cada uno de estos algoritmos con los parámetros por defecto. Es importante mencionar que para todos los experimentos se empleó la estrategia de validación cruzada a diez pliegues.

#### 4.4. Configuración experimental y resultados

Como se ha mencionado en secciones anteriores, el objetivo principal de este trabajo fue determinar hasta qué punto es posible construir un clasificador, mediante el enfoque tradicional de clasificación de textos, capaz de determinar cuándo un tuit es importante o no dentro de un dominio de interés particular para un analista de reputación en línea, con la utilización exclusiva de las palabras contenidas en el mismo tuit.

Para lograr dicho objetivo se propusieron dos conjuntos de experimentos, mismos que se describen a continuación:

**Experimento 1** . Evaluar el impacto que tiene emplear palabras simples como forma de representación, *i.e.*, unigramas de palabras, en la clasificación de tuits importantes. La hipótesis principal de este experimento sugiere que existen algunos términos característicos del dominio de interés (Automotriz, Banca, Musical y Universidades) que permiten identificar de manera eficiente cuándo un tuit es importante.

**Experimento 2** . Evaluar el impacto que tienen las secuencias de palabras de longitud dos, *i.e.*, bigramas de palabras, en la clasificación de tuits importantes. La hipótesis principal de este experimento sugiere que para determinados dominios existen secuencias de palabras, frecuentemente utilizadas por los usuarios, las cuáles permitirían a un ARL identificar los tuits importantes.

Finalmente, para ambos experimentos se emplearon dos diferentes esquemas de pesado: *booleano* y *Tf-Idf*. El objetivo de emplear ambos esquemas de pesado fue valorar el impacto que tiene la presencia de ciertos términos contra la frecuencia de aparición de los mismos.

Los resultados de los experimentos planteados pueden observarse en la Tabla 2 y en la Tabla 3. Observe que en ambas tablas se evaluó el aporte de la forma del esquema de pesado *boolean* y *Tf-Idf* tanto con un clasificador bayesiano (NB) como con las máquinas de vectores de soporte (SVM). Los resultados de los experimentos aparecen en términos de su *Precisión (P)*, *Recuerdo (R)* y de la medida *F* (vea Sección 4.1). Cada línea representa los resultados obtenidos en cada una de las categorías existentes dentro de la colección de tuits proporcionados por el RepLab 2013. Agregado a estos datos, cada tabla contiene en la última fila el promedio de la precisión, recuerdo y medida F obtenidos en las cuatro categorías de tuits empleados durante la experimentación.

Observe que en ambos experimentos, *i.e.*, empleando tanto unigramas (Tabla 2) como bigramas (Tabla 3) de palabras, es claro que la frecuencia de los términos no aporta elementos suficientes al clasificador para determinar cuándo un tuit es o no importante dentro de una determinada categoría. Los experimentos realizados muestran, hasta cierto punto, que la sola presencia de los términos (*i.e.*, ponderado booleano) permite a los algoritmos de clasificación automática contar con mejores elementos para discriminar tuits importantes de los que no lo son.

Note que los resultados obtenidos empleando el clasificador Naïve Bayes (NB) son muy cercanos a los resultados obtenidos al emplear Máquinas de

**Tabla 2.** Resultados de la clasificación empleando unigramas de palabras

| unigramas            |             |             |             |             |             |             |             |             |             |             |             |             |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Categoría de Tuits   | Booleano    |             |             |             |             |             | TF-IDF      |             |             |             |             |             |
|                      | NB          |             |             | SVM         |             |             | NB          |             |             | SMV         |             |             |
|                      | P           | R           | F           | P           | R           | F           | P           | R           | F           | P           | F           |             |
| <i>Automotriz</i>    | 0.61        | 0.62        | <b>0.61</b> | 0.53        | 0.54        | 0.54        | 0.63        | 0.58        | 0.58        | 0.54        | 0.54        | 0.54        |
| <i>Banca</i>         | 0.76        | 0.75        | 0.75        | 0.81        | 0.81        | <b>0.81</b> | 0.29        | 0.72        | 0.69        | 0.79        | 0.80        | 0.79        |
| <i>Musical</i>       | 0.71        | 0.70        | <b>0.70</b> | 0.70        | 0.70        | <b>0.70</b> | 0.65        | 0.64        | 0.64        | 0.67        | 0.66        | 0.66        |
| <i>Universidades</i> | 0.75        | 0.75        | <b>0.75</b> | 0.75        | 0.76        | <b>0.75</b> | 0.75        | 0.72        | 0.72        | 0.70        | 0.71        | 0.70        |
| <i>Promedio</i>      | <i>0.70</i> | <i>0.70</i> | <i>0.70</i> | <i>0.70</i> | <i>0.70</i> | <i>0.70</i> | <i>0.58</i> | <i>0.67</i> | <i>0.66</i> | <i>0.67</i> | <i>0.68</i> | <i>0.67</i> |

**Tabla 3.** Resultados de la clasificación empleando bigramas de palabras

| bigramas             |             |             |             |             |             |             |             |             |             |             |             |             |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Categoría de Tuits   | Booleano    |             |             |             |             |             | TF-IDF      |             |             |             |             |             |
|                      | NB          |             |             | SVM         |             |             | NB          |             |             | SMV         |             |             |
|                      | P           | R           | F           | P           | R           | F           | P           | R           | F           | P           | F           |             |
| <i>Automotriz</i>    | 0.52        | 0.55        | 0.52        | 0.50        | 0.54        | <b>0.55</b> | 0.63        | 0.53        | 0.49        | 0.52        | 0.54        | 0.52        |
| <i>Banca</i>         | 0.74        | 0.75        | 0.75        | 0.80        | 0.80        | <b>0.79</b> | 0.74        | 0.71        | 0.71        | 0.79        | 0.79        | 0.77        |
| <i>Musical</i>       | 0.68        | 0.68        | 0.68        | 0.70        | 0.70        | <b>0.69</b> | 0.64        | 0.63        | 0.63        | 0.69        | 0.69        | 0.68        |
| <i>Universidades</i> | 0.72        | 0.72        | <b>0.72</b> | 0.72        | 0.72        | 0.69        | 0.76        | 0.65        | 0.65        | 0.69        | 0.69        | 0.64        |
| <i>Promedio</i>      | <i>0.67</i> | <i>0.68</i> | <i>0.68</i> | <i>0.68</i> | <i>0.69</i> | <i>0.68</i> | <i>0.69</i> | <i>0.63</i> | <i>0.62</i> | <i>0.68</i> | <i>0.66</i> | <i>0.65</i> |

Vectores de Soporte (SVM), lo cual es un indicador de que los resultados del método propuesto no dependen del algoritmo de clasificación. Esta situación es en realidad un indicativo de que la forma de representación propuesta es apropiada para el problema planteado, es decir, el uso de palabras (unigramas y bigramas) aportan elementos suficientes para que un algoritmo de clasificación automática sea capaz de distinguir tuits importantes de aquellos de no lo son.

En la Figura 1 se muestra de forma gráfica y resumida el comportamiento de los diferentes experimentos realizados para cada una de las cuatro categorías del conjunto de datos empleado. Los resultados se muestran en términos de la medida  $F$ .

De la Figura 1 se pueden extraer varias conclusiones preliminares. En primer lugar parece ser claro que el uso de bigramas de palabras tienen un desempeño menor que con los unigramas de palabras. Este comportamiento queda reflejado también en las filas de *Promedio* en las Tablas 2 y 3 respectivamente, donde se puede notar, bajo una configuración con esquema de pesado *booleano*, un decremento en la medida  $F$ , de un valor de 0.70 usando unigramas a un valor de 0.68 usando bigramas. Sin embargo, a pesar de que el desempeño es relativamente menor, no es posible asegurar que los bigramas de palabras no aportan información importante debido a que las diferencias en el desempeño no parecieran ser significativas. Más experimentos necesitan ser realizados en este punto para descartar la hipótesis planteada en nuestro experimento 2.

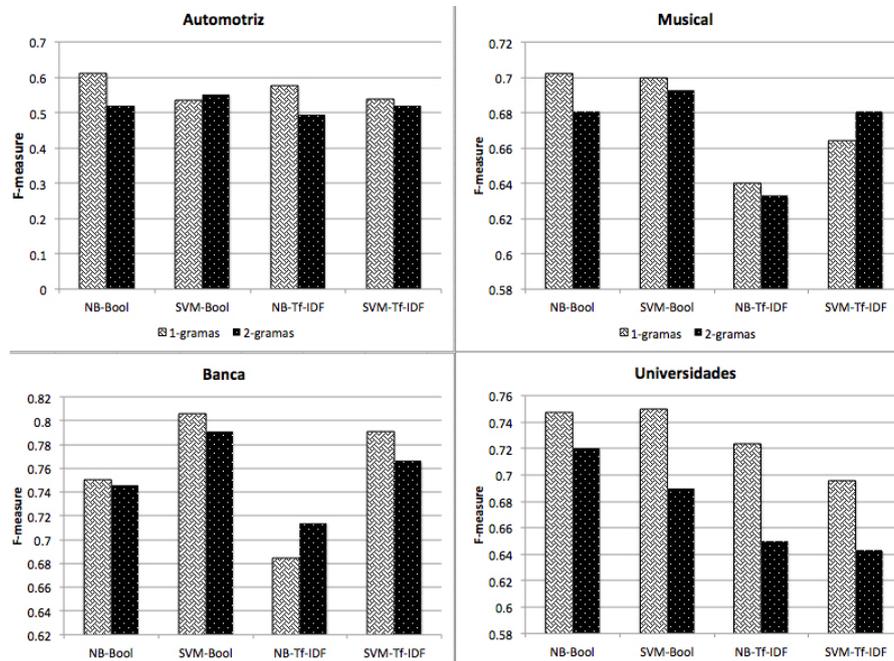


Fig. 1. Comparación de los resultados obtenidos en términos de la medida  $F$

Agregado a esto, es importante resaltar que los mejores resultados de la clasificación se obtienen en la categoría *Banca*, lo cual indica hasta cierto punto, que los usuarios tienden a usar el mismo vocabulario cuando expresan opiniones sobre instituciones bancarias, permitiendo esto que un clasificador sea capaz de aprender cuando la ocurrencia de determinados términos está asociada a la importancia de los tuits. Por el contrario, el dominio *Automotriz* fue en el que los resultados de clasificación resultaron ser más bajos.

Los resultados obtenidos en el dominio *Automotriz* nos hace creer que el tipo de lenguaje empleado entre los usuarios es más variado en este dominio en específico al momento de referir sus opiniones, como puede verse en la Tabla 4 donde se presentan los bigramas con mayor ganancia de información para dos dominios, en particular, para el dominio *Musical* y *Automotriz*. Es claro que para el caso del dominio *Automotriz* las opiniones no mencionan directamente la entidad de la que se opina, algunos de sus términos bien pueden representar al dominio *Banca* o *Universidad*, lo que puede ser la causa del desempeño pobre en este dominio.

En general, de los experimentos podemos afirmar que la presencia de ciertos términos, en particular palabras, aporta elementos suficientes que permiten distinguir tuits importantes dentro de una determinada de categoría de tuits.

**Tabla 4.** Ejemplos de bigramas con mayor ganancia de información para los dominios *Musical* y *Automotriz*

|                   |   |
|-------------------|---|
| <i>Musical</i>    | jennifer lopez, ac dc, alicia key, a shakira, de jennifer, via youtube, shakira y, bon jovi, que shakira, de maroon, por cantar, millones de              |
| <i>Automotriz</i> | anuncio gratis, download free, pdf download, manual pdf, gratis yamaha, de robar, favor rt, acaban de, de honda, con yamaha, de wolswagen, service manuav |

## 5. Conclusiones y trabajo futuro

Se presentó un método de clasificación de textos para la detección de prioridad de tuits. Se encontró que efectivamente existen palabras que son comúnmente utilizadas por los usuarios para referir una opinión, mismas que permiten a un sistema de clasificación automática determinar cuándo un tuit es o no importante para los propósitos de un Analista de Reputación en Línea.

Se realizaron varios experimentos los cuales fueron pensados de tal forma que nos permitieran validar varias hipótesis, las cuales se listan a continuación: *i)* se mostró que la presencia de los términos es suficiente para que un clasificador aprenda a distinguir tuits importantes de aquellos que no lo son; *ii)* se comprobó que la forma de representación propuesta es la que permite resolver el problema con cierta efectividad, y que los resultados no dependen del algoritmo de aprendizaje; *iii)* finalmente, se pudo observar que el uso de bigramas de palabras no son lo suficientemente descriptivos como para superar el desempeño obtenido por una representación basada en unigramas. Se concluyó que, debido a las diferencias poco significativas en el desempeño, más experimentos hacen falta para aceptar o descartar el aporte de los bigramas en la tarea propuesta.

Como trabajo futuro se planea realizar un análisis más detallado de la relevancia de los atributos propuestos en este trabajo, es decir, evaluar el aporte tanto de los unigramas como de los bigramas en el problema atacado. Para esto se pretende realizar más experimentos aplicando estrategias de selección de atributos, los cuales nos permitirán identificar al verdadero valor discriminativo de los atributos. Agregado a esto, se evaluará el método propuesto en el subconjunto de tuits en idioma Inglés del RepLab.

Los resultados obtenidos muestran que el uso de una representación basada en las palabras contenidas en los tuits puede ser exitosa, pues obtuvo resultados competitivos a los reportados en la competencia de RepLab 2013, al mismo tiempo queda claro que aún existe un amplio margen de trabajo que resta por hacer. Es importante mencionar que de entre los equipos participantes en la tarea del RepLab 2013, el desempeño del sistema mejor evaluado fue de un  $F=0.63$  en promedio, mientras que el método base obtuvo un  $F=0.60$  [1].

**Agradecimientos.** Agradecemos el apoyo otorgado por la Universidad Autónoma Metropolitana Unidad Cuajimalpa y el SNI-CONACyT.

## Referencias

1. Amigó, E., de Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín-Wanton, T., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2013: Evaluating online reputation monitoring systems. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF. Lecture Notes in Computer Science, vol. 8138, pp. 333–352. Springer (2013)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)
3. Balahur, A., Tanev, H.: Detecting entity-related events and sentiments from tweets using multilingual resources. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF (Online Working Notes/Labs/Workshop) (2012)
4. Duda, R., Hart, P.: Pattern classification and scene analysis. Wiley (1996)
5. Garner, S.R.: Weka: The Waikato environment for knowledge analysis. In: Proc. of the New Zealand Computer Science Research Students Conference. pp. 57–64 (1995)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2009)
7. Jean-Valère Cossu, Benjamin Bigot, L.B.M.M.X.B.G.S.R.D.V.B.J.M.T.M., El-Bèze, M.: Lia@replab 2013. In: CLEF 2013 Eval. Labs and Workshop Online Working Notes (2013)
8. Leon-Martagón, G., Villatoro-Tello, E., Jiménez-Salazar, H., Sánchez-Sánchez, C.: Análisis de polaridad en twitter. Journal of Research in Computing Science. 62, 69–78 (2013)
9. LiMoSINe: Replab 2013 – an evaluation campaign for online reputation management systems (2014), Última visita 15 de marzo de 2014
10. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24(5), 513–523 (Aug 1988)
11. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
12. Spina, D., Carrillo de Albornoz, J., Martín, T., Amigó, E., Gonzalo, J., Giner, F.: UNED Online Reputation Monitoring Team at RepLab 2013. In: CLEF 2013 Eval. Labs and Workshop Online Working Notes. CLEF (2013)
13. Twitter, I.: About twitter (2014), Última visita 15 de marzo de 2014