UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
POSTGRADO EN MODELOS ALEATORIOS

LAS VOCES DE TWITTER SOBRE LAS DIMENSIONES DE POBREZA EN VENEZUELA: UN ENFOQUE DE BIGDATA.

Autor: Lic. Manuel Solorzano.
Tutor: Dr. Daniel Barraéz.

Trabajo de Grado de Maestría presentado ante la ilustre Universidad Central de Venezuela para optar al título de Magister Scientiarium en Modelos Aleatorios.

Caracas, Venezuela
Octubre 2018

Dedicatoria

A mis padres, mis familiares y amigos más cercanos.

Agradecimiento

Primeramente, Dios todo poderoso. Mis padres, especialmente a mi papá. Mi tutor por darme la opurtunidad.

Índice general

Veredicto

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
COMISIÓN DE ESTUDIOS DE POSTGRADO

Comisión de Estudios de Postgrado

VEREDICTO

Quienes suscriben, miembros del jurado designado por el Consejo de la Facultad de Ciencias de la Universidad Central de Venezuela, para examinar el **Trabajo de Grado**presentado por: Manuel Javier Solórzano Torres**, Cédula de identidad** 19.224.554, bajo el título "**LAS VOCES DE TWITTER SOBRE LAS DIMNESIONES DE POBREZA EN VENEZUELA: UN ENFOQUE DE BIGDATA.**", a fin de cumplir con el requisito legal para optar al grado académico de **MAGÍSTER SCIENTIARUM, MENCIÓN MODELOS ALEATORIOS**, dejan constancia de lo siguiente:

1.- Leído como fue dicho trabajo por cada uno de los miembros del jurado, se fijó el día 14 de diciembre de 2018 a las 10:30am., para que el autor lo defendiera en forma pública, lo que éste hizo en LAPROESA, mediante un resumen oral de su contenido, luego de lo cual respondió satisfactoriamente a las preguntas que le fueron formuladas por el jurado, todo ello conforme con lo dispuesto en el Reglamento de Estudios de Postgrado.

2.- Finalizada la defensa del trabajo, el jurado decidió **Aprobarla** por considerar, sin hacerse solidario con la ideas expuestas por el autor, que se ajusta a lo dispuesto y exigido en el Reglamento de Estudios de Postgrado.

Para dar este veredicto, el jurado estimó que el trabajo procesa y analiza información de gran interés y actualidad acerca de la discusión de la pobreza en Venezuela y sus dimensiones, en la red social twitter con herramientas estadísticas avanzadas, de lo cual se levanta la presente ACTA, a los 14 días del mes de diciembre del año 2018, conforme a lo dispuesto en el Reglamento de Estudios de Postgrado, actuó como Coordinador del jurado Daniel Barráez.

Mairene Colina / C.I.12.761.954
Institución UCV

Ricardo Ríos / C.I. 3.949.476
Institución UCV

Daniel Barráez / C.I. 6.197.305
Institución UCV
Tutor

Resumen

La mayoría de los países del mundo definen la pobreza como la falta de dinero. Sin embargo, los propios pobres consideran que su experiencia de la pobreza es mucho más amplia que la carencia de ingresos. Una persona que es pobre puede sufrir múltiples desventajas al mismo tiempo. Enfocarse en un solo factor, tal como el ingreso, no es suficiente para capturar la verdadera realidad de la pobreza.

La noción de pobreza multidimensional pretende agrupar múltiples desventajas que presencian los pobres. Se entiende por Índice de Pobreza Multidimensional (IPM), como el índice que identifica múltiples carencias a nivel de los hogares y las personas en los ámbitos de salud, la educación y las condiciones de vida.

En este proyecto análizaremos tweets relacionados con la palabra pobreza y sus dimensiones mediante la aplicación web de Twitter, para realizar minería de opinión e identificar las opiniones expresadas de los usuarios referente a pobreza multidimensional, plantear soluciones y recomendaciones que permita abordar el tema desde una perspectiva política y social. Empleamos técnicas de análisis de sentimiento y la asignación latente de Dirichlet.

Palabras Claves: Twitter, tweets, pobreza multidimensional, análisis de sentimiento, asignación latente de Dirichlet, ingreso, Venezuela, minería de opinión, salud, educación geolocalización.

Abstract

Most countries in the world define poverty as the lack of money. However, the poor themselves consider that their experience of poverty is much broader than the lack of income. A person who is poor can suffer multiple disadvantages at the same time. Focusing on a single factor, such as income, is not enough to capture the true reality of poverty.

The notion of multidimensional poverty seeks to group multiple disadvantages that the poor see. The Multidimensional Poverty Index (MPI) is understood as the index that identifies multiple deficiencies at the level of households and the people in the areas of health, education and standard of living.

In this project we will analyze tweets related to the word poverty and its dimensions through the Twitter web application, to perform opinion mining and identify the opinions expressed by users regarding multidimensional poverty, propose solutions and recommendations that allow us to approach the issue from a perspective political and social. We use feeling analysis techniques and the latent Dirichlet allocation.

Key words: Twitter, tweets, multidimensional poverty, sentiment analysis, Latent Dirichlet Allocation, income, Venezuela, opinion mining, health, geolocation, education.

Introducción

En el mundo en el que vivimos, las personas se ven en la necesidad de hacer uso de los sistemas de información. Por ejemplo, en la Web, las personas por lo general publican fotos de vivencias, suben propagandas de sus negocios, publican logros, expresan opiniones, etc. Específicamente en la Red social Twitter los usuarios expresan emociones, pasiones, molestias, desconformidad, necesidad, alegrías, entre otras cosas. Ello ha motivado a una variedad de empresas e investigadores al desarrollo de habilidades en el manejo de opinión a través de esta red social, para tener una mejor visión del mercado y saber dónde (y a quien) dirigir sus productos con mayor éxito de aceptación por parte de los consumidores. En los últimos años Twitter se ha convertido en una especie de periódico digital, con muchas opiniones sociales y políticas.

Este proyecto tiene como objetivo realizar un seguimiento en Twitter, para explorar las principales opiniones y/o emociones de los usuarios de Twitter en Venezuela, en temas de pobreza multidimensional, como nueva fuente de datos e información. En el capítulo 1, introduciremos la definición de pobreza multidimensional.

Es de vital importancia resaltar que este proyecto no es un estudio basado en encuesta, por lo tanto, la idea de muestra representativa de la población no está contemplada, sino un análisis exploratorio de las opiniones de los usuarios en Twitter en temas de pobreza. Como es bien conocido por los expertos y las personas interesadas y/o relacionadas en el tema, para realizar encuestas se necesita inversión de tiempo (1 mes, 2 meses, todo depende del estudio) en la obtención de los datos, análisis y posteriormente medir problemáticas para la toma de decisiones. Cabe destacar que las encuestas no dejan de ser una forma de medir las principales problemáticas sociales. Por otro lado, se podría inferir que la red social Twitter no es usado por la clase social más pobre y vulnerable. Sin embargo, esta herramienta es útil para captar opiniones sociales de los usuarios. Además, Twitter permiten obtener análisis geolocalizados en poco tiempo.

En Venezuela, basta mirar las noticias unos poco minutos para encontrar artículos referentes a seguridad, educación, salud, hambre, salarios entre otros, no muy favorables para los ciudadanos.

En este proyecto análizaremos tweets relacionados con la palabra pobreza y sus dimensiones mediante la aplicación web de Twitter, para realizar minería de opinión e identificar las opiniones expresadas de los usuarios referente a pobreza multidimensional, plantear soluciones y recomendaciones que permita abordar el tema desde una perspectiva política y social. Empleamos técnicas de análisis de sentimiento y la asignación latente de Dirichlet.

Hemos desarrollado los cápitulos:

- Pobreza Multidimensional,
- Aprendizaje Automático,
- Análisis de Sentimiento, y
- Resultados y Análisis

Capítulo 1

Pobreza Multidimensional

La mayoría de los países del mundo definen la pobreza como la falta de dinero. Sin embargo, los propios pobres consideran que su experiencia de la pobreza es mucho más amplia que la carencia de ingresos. Una persona que es pobre puede sufrir múltiples desventajas al mismo tiempo - por ejemplo, puede tener una mala salud o estar desnutrida, puede carecer de agua limpia o electricidad, tener un trabajo precario o tener un bajo nivel educativo. Enfocarse en un solo factor, tal como el ingreso, no es suficiente para capturar la verdadera realidad de la pobreza.

1. Índice de Probreza Multidimensional

El índice de pobreza multidimensional (IPM) o índice multidimensional de pobreza (IMP), en inglés: multidimensional poverty index o MPI, es un índice de pobreza estadístico sobre la situación de las personas por paises, elaborado desde 2010.
El IPM es elabarado por el **Programa de las Naciones Unidas para el Desarrollo** (PNUD) de la **Organización de las Naciones Unidas** (ONU) en colaboración con la OPHI (Oxford Poverty & Human Development Initiative, Iniciativa de pobreza y Darrollo Humano de Oxford), y se presentó en el aniversario del "Informe Anual Mundial sobre el Desarrollo Humano" del PUND.
Se entiende por Índice de Pobreza Multidimensional (IPM), como el índice que identifica múltiples carencias a nivel de los hogares y las personas en los ámbitos de salud, la educación y las condiciones de vida. Cada miembro de una familia es clasificado como pobre o no pobre en función del número de carencias que experimente su hogar. El IPM ofrece un valioso complemento a las herramientas de medición de la pobreza basadas en los ingresos.

2. Metodolagía

A cada persona se le asigna un puntaje de carencia de acuerdo a su o las carencias de su hogar en cada uno de los 10 indicadores. La puntuación máxima de carencias es del 100 %, con cada dimensión igualmente ponderada; por lo tanto, la privación máxima en cada dimensión es del 33,3 %. La dimensiones educación y salud tienen dos indicadores cada una, de modo que cada indicador vale 33.3/2, o 16.7 %. La dimensión condiciones de vida tiene seis indicadores, por lo que cada indicador vale 33.3/6, o 5.6 %.
Los umbrales de los indicadores para los hogares a ser considerados como carencias son los siguientes:

- **Educación:**
  1. Años de escolaridad: ningún miembro del hogar ha completado seis años de escolaridad.
  2. Asistencia escolar: un niño en edad escolar (hasta octavo grado) no asiste a la escuela. [1]
- **Salud:**
  3. Nutrición: un miembro del hogar (para quien existe información nutricional) está desnutrido, medido por el índice de masa corporal para adultos (mujeres de 15 a 49 años en la mayoría de las encuestas) y por el z-score de altura para la edad (en inglés: height-for-age z-score) calculado basado en las normas de la Organización Mundial de la Salud para niños menores de 5 años.
  4. Mortalidad infantil: un niño ha muerto en el hogar dentro de los cinco años anteriores a la encuesta. [2]
- **Condiciones de Vida:**
  5. Electricidad: no tener acceso a la electricidad.
  6. Agua potable: no tener acceso a agua potable o tener acceso a agua potable a través de una fuente que se encuentra a 30 minutos o más caminando.

[1] Se permite hasta un año de inscripción tardía en la escuela primaria para evitar el conteo de desajustes entre el cumpleaños y el comienzo del año escolar como una privación.

[2] Algunas encuestas no recopilan información sobre el momento en que ocurrió la muerte de un niño; en tales casos, se cuenta cualquier muerte infantil notificada por una madre de 35 años o menos.

7. Saneamiento: el hogar no tiene un baño con condiciones suficientes o si su baño es compartido. [3]
8. Combustible para cocinar: utiliza combustible para cocinar "sucio" (estiércol, medera o carbón vegetal).
9. Suelo: el piso del hogar tiene suciedad, es de arena, tierra o estiércol.
10. Bienes: no tener al menos un activo relacionado con la información (radio, televisión o teléfono [4]) o que tengan al menos un activo relacionado con la información pero que no tenga al menos un activo relacionado con la movilidad (bicicleta, motocicleta, coche, camión, carro de animales o lancha a motor) o al menos un bien relacionado con el sustento (refrigerador, tierra cultivable [5] o ganado [6]).

3. ¿Para qué sirve el IPM?

Un Índice de Pobreza Multidimensional (IPM) se puede utilizar para implementar políticas públicas con sustento empírico al elaborar programas sociales más rentables que apunten a las necesidades de las personas viviendo en la pobreza. La estructura de la metodología Alkire Foster [7] tiene propiedades que hacen que un IPM sea particularmente útil para informar de manera transparente a la política pública. Entre otras cosas, pueden utilizarse para:

- Producir medidas oficiales de pobreza multidimensional.
- Comparar la incidencia y la intensidad de la pobreza entre países.
- Comparar grupos subnacionales, como regiones, poblaciones urbanas / rurales y grupos étnicos.
- Comparar la composición de la pobreza por dimensiones e indicadores.
- Informar sobre los cambios en la pobreza a lo largo del tiempo.

[3] El agua potable y el saneamiento mejorado se definen en los Objetivos de Desarrollo del Milenio.

[4] Incluyendo teléfonos fijos y móviles.

[5] Cualquier tamaño de tierra utilizable para la agricultura.

[6] Un caballo, una cabeza de ganado, dos cabras, dos ovejas o 10 gallinas.

[7] El Método Alkire-Foster (AF) es una forma de medir la pobreza multidimensional desarrollada por Sabina Alkire y James Foster de OPHI. Sobre la base de las medidas de pobreza de Foster-Greer-Thorbecke, se trata de contar los diferentes tipos de privación que las personas experimentan al mismo tiempo, como la falta de educación o empleo, o los bajos niveles de salud o de vida.

Para identificar a los pobres multidimensionales, las puntuaciones de privación para cada indicador se suman para obtener la puntuación de privación del hogar. Se utiliza un valor de corte del 33,3 por ciento, que equivale a 1/3 de los indicadores ponderados, para distinguir entre los pobres y no pobre. Si el puntaje de privación es de 33.3 por ciento o más, ese hogar (y todos los que están en él) es multidimensionalmente pobre. Los hogares con un puntaje de privación del 20 por ciento o más alto pero menos del 33.3 por ciento están cerca de la pobreza multidimensional. Los hogares con un puntaje de privación del 50 por ciento o más son extremadamente pobres multidimensionalmente.

Particularmente, en este proyecto definimos diccionarios de palabras para identificar mediante el contenido de los tweets en que dimensión de pobreza se clasifica el tweet. Similarmente, se ajustó un diccionario para los ingresos percibidos. Con estos diccionarios se realizó un filtro para relacionar los textos emitidos por los usuarios en los tweets, con las dimensiones de pobreza e ingresos. En el filtrado de los texto hemos utilizado la librería *(regular expressions, re)* del lenguaje de programación **Python.** Finalmente, haciendo uso de estas clasificaciones mediante la ayuda de estos diccionarios queremos identificar las opiniones expresadas por los usuarios en temas pobreza.

Los Cuadros 1 y 2. contienen los diccionarios implementados en este proyecto para clasificar los tweets en dimensión de pobreza e ingresos.

 **Dimensión Condiciones de vida**				
Servicio	Corte	Desechos	Carbón	Agua
Rancho	Vivir	Transporte	Piso	Metro
Comida	Vivienda	Leña	Habitación	Duermen
Apagón	Terreno	Propiedad	Excretas	Bicicleta
Pasajeros	Casa	Moto	Recolecció	Cloacas
Pasaje	Apartamento	Alquiler	Residuos	Televisión
Aseo	Hogar	Teléfono	Bienes	Baños
Luz	Techo	Carro	Potable	Radio
Electricidad	Calles	Fogón	Desperdicio	metro bus

CUADRO 1. Diccionario de palabras dimensión condiciones de vida.

 **Dimensión Salud**		 **Dimensión Educación**		 **Ingresos Percibidos**	
Medicina	Moribundo	Educación	Educativo	Trabajo	Ingresos
Medicamentos	Muerte	Escolar	Educar	Trabajadores	Salario
Basura	Saludable	Escuela	Estudiantado	Empresas	Salarios
Médicos	Comer	Preescolar	Bachillerato	Empresario	Quincena
Salud	Comida	Liceo		Chamba	Sueldos
Hospitalizado	Salud	Colegio		Empleo	Sueldo
Enfermedad	Hambre	Estudiante		Empleado	Pago
Enfermeras	Desnutrido	Estudiantil		Puesto	Pensiones
Enfermo	Nutrición	Universidad		Vacante	Obrero
Farmacia	Escasez	Clases		Sindicato	Ingreso
Hospital	Alimentos	Deserción		Dólar	

CUADRO 2. Diccionario de palabras dimensión salud, dimensión educación e ingresos.

Capítulo 2

Aprendizaje Automático

Cuando la supercomputadora **Deep Blue** de **IBM** venció al campeón de ajedrez **Gary Kasparov** en 1997, muchos se sorprendieron ante el poder de estas máquinas.

Dos décadas después, la inteligencia artificial está presente en la banca, la medicina y en programas populares como los predictores de palabras de los celulares.

En 2016, AlphaGo, un programa informático de la filial de Google **Deep Mind,** ganó un duelo con el campeón del mundo del juego Go.

Y en 2017 el Instituto Tecnológico de Massachussetts (MIT) anunció que su algoritmo **DeepMoji,** puede analizar emojis para detectar el sarcasmo en Twitter.

La inteligencia artificial está en su mejor época, pues la gran mayoría de tecnologías emergentes poseen módulos que les permiten interactuar de forma más autónoma y humanizada. Dentro de las múltiples áreas que comprende la inteligencia artificial, se encuentra la de aprendizaje automático, en algunos textos se le conoce como: aprendizaje de máquina o Machine Learning (ML) por su traducción en inglés.

El aprendizaje automático es una técnica de análisis de datos que enseña a los ordenadores a hacer lo que resulta natural para las personas y los animales: **aprender de la experiencia.** Los algoritmos de aprendizaje automático emplean métodos de cálculo para "aprender" información directamente de los datos sin depender de una ecuación predeterminada como modelo. Los algoritmos mejoran su rendimiento de forma adaptativa a medida que aumenta el número de muestras disponibles para el aprendizaje. [2]

El aprendizaje automático se ha convertido en una técnica clave para resolver problemas en áreas tales como:

- **Finanzas computacionales:** Para la calificación crediticia y el trading algorítmico.
- **Procesamiento de imágenes y visión artificial:** Para el reconocimiento facial, la detección de movimiento y la detección de objetos.

- **Biología computacional:** Para la detección de tumores, el descubrimiento de fármacos y la secuenciación del ADN.
- **Producción de energía:** Para la previsión de la carga y el precio.
- **Automoción, sector aeroespacial y fabricación:** Para el mantenimiento predictivo.
- **Procesamiento del lenguaje natural:** Para aplicaciones de reconocimiento de voz.
- **Análisis de texto:** Para aplicaciones de la mineria de opinión.

Algunos sistemas de Aprendizaje Automático intentan eliminar toda necesidad de intuición o conocimiento experto de los procesos de análisis de datos, mientras otros tratan de establecer un marco de colaboración entre el experto y la computadora. De todas formas, la intuición humana no puede ser reemplazada en su totalidad, ya que el diseñador del sistema ha de especificar la forma de representación de los datos y los métodos de manipulación y caracterización de los mismos. Dependiendo de las necesidades del problema, el ambiente en el que se van a desenvolver y los factores que afectarán la toma de decisiones, podemos encontrar distintos tipos de algoritmos de aprendizaje, tales como: supervisado, no supervisado y por refuerzo.

En este proyecto expondremos brevemente en que consiste el aprendizaje automático supervisado y aprendizaje automático no supervisado.

1. Tipos de Aprendizaje Automático.

1.1. Aprendizaje Supervisado (Supervised machine learning).

Un algoritmo de aprendizaje supervisado toma un conjunto conocido de datos de entrada y respuestas conocidas para estos datos (salidas) y entrena un modelo con objeto de generar predicciones razonables como respuesta a datos nuevos.

La palabra clave “supervisado” viene de la idea de tener un conjunto de datos previamente etiquetado y clasificado, es decir, tener un conjunto de muestra el cual ya se sabe a qué grupo, valor o categoría pertenecen los ejemplos. Con este grupo de datos, el cual llamamos datos de entrenamiento, se realiza el ajuste al modelo inicial planteado. De esta forma es como el algoritmo va “aprendiendo” a clasificar las muestras de entrada comparando el resultado del modelo, y la etiqueta real de la muestra, realizando las compensaciones

respectivas al modelo de acuerdo a cada error en la estimación del resultado. Por ejemplo, el aprendizaje supervisado ha sido utilizado para la programación de vehículos autónomos[3]. Algunos métodos y algoritmos que podemos implementar son los siguientes:

- Clasificación:
  - K vecinos más cercanos (K-nearest neighbors, KNN)
  - Máquinas de vectores de soporte (Support vector machines, SVM)
  - Clasificador Bayesiano ingenuo (Naïve Bayes classifier)
  - Análisis Discriminante (Discriminant Analysis)
- Regresión:
  - Redes neuronales artificiales (Artificial neural networks)
  - Árboles de decisión (Decision trees)
  - Regresión logística, Regresión Lineal (Logistic regression, Linear Regression)

1.2. Aprendizaje No Supervisado (Unsupervised machine learning).

Todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos.

Es decir, a diferencia del supervisado, los datos de entrada no están clasificados ni etiquetados, y no son necesarias estas características para entrenar el modelo. Dentro de este tipo de algoritmos, el agrupamiento o clustering en inglés, es el más utilizado, ya que particiona los datos en grupos que posean características similares entre sí. Una aplicación de estos métodos es la compresión de imágenes[3]. Entre los principales algoritmos de tipo no supervisado destacan:

- K-medias, k-Medoids, c-medias difusas (K-Means, k-Medoids, Fuzzy c-Means)
- Agrupamiento jerárquico (Hierarchical clustering)
- Redes Neuronales (Neural Networks)
- Asignación Latente de Dirichlet (Latent Dirichlet Allocation, LDA)

2. Asignación Latente de Dirichlet.

Para fines de este proyecto vamos a considerar la Asignación Latente de Dirichlet (en inglés, Latent Dirichlet Allocation, LDA), que es un algoritmo no supervisado que pretende describir un conjunto de observaciones (acá las observaciones son documento) como una combinación de categorías diferentes. Las características son la presencia (o frecuencias de presencia) de cada palabra y las categorías son los temas. Los temas aprenden con una distribución de probabilidad a través de las palabras que se generan en cada documento. Cada documento, a su vez, se describe como una combinación de temas. La idea es generar temas, aunque algunas veces similares, pero con significado diferentes. Veamos algunas definiciones necesarias.

- Una **palabra** es la unidad básica de datos discretos, definida como un elemento de un vocabulario indexado por $\{1, \ldots, V\}$. Representamos palabras utilizando vectores de base unitaria que tienen un solo componente igual a uno y todos los demás componentes son iguales a cero. Por lo tanto, al utilizar superíndices para denotar componentes, la n-enésima palabra en el vocabulario se representa mediante un $V-$vector w tal que $w^v = 1$ y $w^u = 0$ para $u \neq v$. En este proyecto denotaremos las palabras como términos.
- Un **documento** es una secuencia de N palabras denotadas por $w = (w_1, w_2, \ldots, w_N)$, donde w_N es el n-enésima palabra en la secuencia. Para fines de este proyecto los documentos son los tweets.
- Un **corpus** es una colección de documentos M denotados por $D = \{w_1, w_2, \ldots, w_M\}$.

Formalmente, LDA es un modelo probabilístico generativo de un corpus. La idea básica es que los documentos se representan como mezclas aleatorias sobre temas latentes, donde cada tema se caracteriza por una distribución sobre palabras[1]

LDA asume el siguiente proceso generativo para cada documento **w** en un corpus D:

(1) Elejir N $\sim Poisson(\xi)$.
(2) Elejir $\theta \sim Dir(\alpha)$.
(3) Para cada una de las N palabras w_n :

[1]Nos referimos a las variables multinomiales latentes en el modelo LDA como temas para explotar las intuiciones orientadas al texto, pero no hacemos afirmaciones epistemológicas con respecto a estas variables latentes más allá de su utilidad para representar distribuciones de probabilidad en conjuntos de palabras.

(a) Elejir un tópico $z_n \sim Multinomial(\theta)$.

(b) Elejir una palabra w_n de $p(w_n|z_n, \beta)$, una probabilidad multinomial condicionada al tópico z_n.

Se hacen varias suposiciones simplificadoras en este modelo básico. Primero, la dimensionalidad k de la distribución de Dirichlet (y por lo tanto la dimensionalidad de la variable temática z) se asume conocida y fija. Segundo, las probabilidades de palabra están parametrizadas por una matriz β, de tamaño $k \times V$, donde $\beta_{ij} = p(w^j = 1|z^i = 1)$, que por ahora tratamos como una cantidad fija que debe ser estimada. Finalmente, la suposición de Poisson no es crítica para nada de lo que sigue y se pueden usar distribuciones de longitud de documentos más realistas según sea necesario. Además, tenga en cuenta que N es independiente de todas las demás variables generadoras de datos (θ y z). Por lo tanto, es una variable auxiliar y generalmente ignoraremos su aleatoriedad en el desarrollo posterior.

Una variable aleatoria Dirichlet $k - dimensional$ θ puede tomar valores en $(k-1) - simplex$ (un $k - vector$ θ se encuentra en $(k-1) - simplex$ si $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$), y tiene la siguiente densidad de probabilidad en este símplex:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\Pi_{i=1}^{k} \Gamma\left(\alpha_i\right)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \tag{2.1}$$

donde el parámetro α es un $k - vector$ con componentes $\alpha_i > 0$, y donde $\Gamma(x)$ es la función Gamma. La distribución Dirichlet es conveniente en el símplex, está en la familia exponencial, tiene estadísticas suficientes de dimensiones finitas y está conjugado con la distribución multinomial. En la Sección 5 de [4], estas propiedades facilitarán el desarrollo de algoritmos de inferencia y estimación de parámetros para LDA.

Dado los parámetros α y β, la distribución conjunta de una mezcla de tópicos θ, un conjunto de N tópicos z, y un conjunto de N palabras w, está dada por:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \tag{2.2}$$

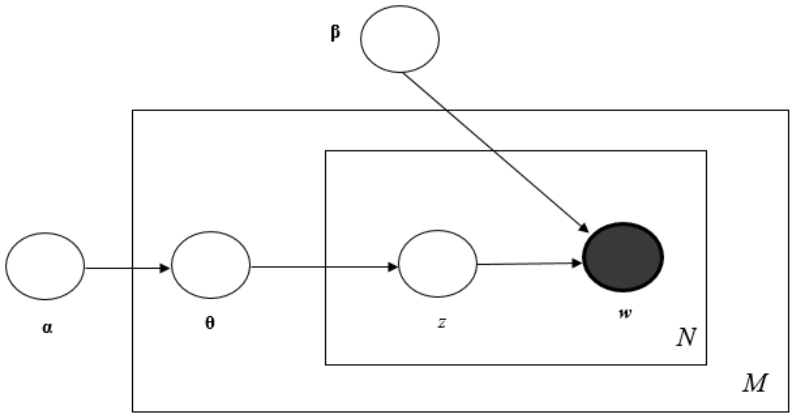


FIGURA 2.1. Representación gráfica de modelo de LDA. Los cuadros son "placas" que representan réplicas. La placa exterior representa documentos, mientras que la placa interior representa la elección repetida de tópicos y palabras dentro de un documento.

donde $p(z_n|\theta)$ es simplemente θ_i para un único i tal que $z_n^i = 1$. Integrando más de θ y sumando z, obtenemos la distribución marginal de un documento:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta. \tag{2.3}$$

Finalmente, tomando el producto de las probabilidades marginales de documentos únicos, obtenemos la probabilidad de un corpus:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d. \tag{2.4}$$

El modelo LDA se representa como un modelo gráfico probabilístico en la Figura 2.1. Como la figura deja en claro, la representación LDA tiene tres niveles. Los parámetros α y β son parámetros de nivel de corpus, se supone que se muestrean una vez en el proceso de generación de un corpus. Las variables α_d son variables de nivel de documento, muestreadas una vez por documento. Finalmente, las variables z_{dn} y w_{dn} son variables de nivel de palabra y se muestrean una vez para cada palabra en cada documento.

Es importante distinguir LDA de un simple modelo de agrupamiento multinomial de Dirichlet. Un modelo de agrupamiento clásico implicaría un modelo de dos niveles en el que se muestrea un Dirichlet una vez para un corpus, una variable de agrupación multinomial se selecciona una vez para cada documento en el corpus, y se selecciona un conjunto de palabras para el documento condicional en el clúster variable. Al igual que con muchos modelos de agrupación, dicho modelo restringe la asociación de un documento a un único tópico. LDA, por otro lado, implica tres niveles, y notablemente el nodo tópico se muestrea repetidamente dentro del documento. Bajo este modelo, los documentos se pueden asociar con múltiples tópicos.

Las estructuras similares a la mostrada en la Figura 2.1. A menudo se estudian en modelos estadísticos bayesianos, donde se los denomina modelos jerárquicos (Gelman y otros, 1995) o, más precisamente, como modelos jerárquicos condicionalmente independientes (Kass y Steffey, 1989). Dichos modelos también suelen denominarse modelos empíricos paramétricos de Bayes, un término que se refiere no solo a una estructura de modelo particular, sino también a los métodos utilizados para estimar parámetros en el modelo (Morris, 1983). En [4], se hace una adaptación a el enfoque empírico de Bayes para estimar parámetros como α y β en implementaciones simples de LDA, también se considera enfoques Bayesianos más completos.

Para hacer uso de esta herremienta, en este proyecto se empleo una librería del paquete **R** llamada *topicmodels.* El paquete topicmodels proporciona una interfaz al código C para los modelos de Asignación Latente de Dirichlet (LDA) y Modelos de Temas Correlacionados (CTM) propuesto por *David M. Blei* y coautores y el código C ++ para ajustar los modelos LDA utilizando el muestreo de Gibbs propuesto por *Xuan-Hieu Phan* y coautores.

Capítulo 3

Análisis de Sentimiento

La demanda de infromación sobre opiniones y sentimientos "Lo que otras personas piensan" siempre ha sido una información importante para la mayoria de nosotros durante el proceso de toma de decisiones. Con la creciente disponibilidad y popularidad de los recursos ricos en opinión, como los sitios de reseñas en línea y el blog personal, surgen nuevas oportunidades y desafíos, ya que la gente ahora puede, y lo hace, utilizar activamente la información tecnologías para buscar y entender las opiniones de los demás[5]. La repentina erupción de actividad en el área de la minería de texto o minería de opiniones y el análisis de sentimientos, que se ocupa del tratamiento computacional de opinión, sentimiento, y subjetividad en el texto, ha ocurrido así, al menos en parte, como una respuesta directa al aumento del interés en los nuevos sistemas que tratan directamnete las opiniones como un objeto de primera clase.

La tarea es técnicamnete desafiante y prácticamnete muy útil. Por ejemplo, las empresas siempre quieren conocer la opinión pública o de los consumidores sobre sus productos y servicios. Clientes potenciales también quieren conocer las opinones de los usuarios existentes antes de utilizar un servicio o comprar un producto. Con el crecimiento explosivo de los medios de comunicación social (es decir, críticas, foros de discusión, blogs y redes sociales) en la Web, individuos y organizaciones en estos medios de comunicación, utilizan cada vez más a la opinión pública para su toma de decisiones. Sin embargo, la búsqueda y el seguimiento de los sitios de opinión en la Wed y la destilación de la información contenida en ellos sigue siendo una tarea formidables debido a la proliferación de sitios diversos.

1. El problema de la minería de opinión

En esta sección, definimos el promblema de minería de opinión, el cual nos permite ver una estructura desde el intimidante texto no estructurado y para proporcionar un marco unificado para la investiación actual. La abstracción consta de dos partes: definición de opinión y resumen de opinión [6], pág. 416.

1.1. Definición de opinión.

Utilizamos el siguiente segmento de reseña en iPhone para presentar el problema (se asocia un número de identificación a cada frase para una referencia fácil): " (1)Compré un iPhone hace unos días. (2) Era un teléfono tan bonito. (3) La pantalla táctil era realmente genial. (4) La calidad de voz también era clara. (5) Sin embargo, mi madre estaba enojada conmigo ya que no se lo dije antes de comprarlo. (6) También pensó que el teléfono era demasiado costoso y quería que lo devolviera a la tienda..."

La pregunta es: ¿Qué queremos extraer o que extrato de esta reseña?. Lo primero que notamos es que hay varias opiniones en esta reseña. Las frases (2), (3) y (4) expresan opiniones positivas, mientras que las frases (5) y (6) expresan opiniones o emociones negativas. Seguidamente, notamos que todas las opiniones tienen algunos objetivos. El objetivo de la opinión en la frase (2) es el iPhone como un todo, y los objetivos de las opiniones en las frases (3) y (4) son "pantalla táctil" y "calidad de voz" del iphone respectivamente. El objetivo de la opinión en la frase (6) es el precio del iPhone, pero el objetivo de la opinión/emoción en la frase (5) es "mi", no iPhone. Finalmente, podemos notar a los titulares de opiniones.

El titular de las opiniones en las oraciones (2),(3) y (4) es el autor de la reseña ("yo"), pero en las oraciones (5) y (6) es "mi madre". Con este ejemplo en mente, ahora definimos formalmente el problema de la mineria de opinión. Comenzaremos con el objetivo de opnión. En general, se pueden expresar opiniones sobre cualquier cosa, por ejemplo, un producto, un servicio, un individuo, una organización, un evento o un tema, por cualquier persona u organización. Utilizamos la entidad para denotar el objeto objetivo que ha sido evaluado. Formalmente, tenemos lo siguiente:

Definición 3.1. (Entidad) Una entidad e es un producto, servicio, persona, evento, organización o tema. Se asocia con un par, e:(T,W), donde T es una jerarquía de componentes (o partes), subcomponentes y así sucesivamente, y W es un conjunto de atributos de e. Cada componente o subcomponente también tiene su propio cunjunto de atributos.

Un ejemplo de una entidad es el siguiente:

Ejemplo 3.2. Una marca particular de teléfono celular es una entidad, por ejemplo iPhone. Tiene un conjunto de componentes, como por ejemplo, la batería y la pantalla, y

también un cunjunto de atributos, por ejemplo, calidad de voz, tamaño y peso. El componete de la batería también tiene su propio conjunto de atributos, por ejemplo, la duración de la bateria, y el tamaño de la batería.

Hay dos tipos principales de opiniones: opiniones regulares y opiniones comparativas.

Una opinión comparativa expresa una relación de similitudes o diferencias entre dos o más entidades, y/o una preferencia del formador de opinión basada en algunos de los aspectos compartidos de las entidades. Una opinión comparativa generalmente se expresa utilizando la forma comparativa o superlativa de un adjetivo o adverbio, aunque no siempre.

Más aun, opinión regular es simplemente un sentimiento, una actitud, una emoción o una valoración positiva o negativa sobre una entidad o de un aspecto de la entidad por parte de un titular de opinión. Lo positivo, negativo y neutral se denominan orientaciones de opinión (también llamadas orientaciones de sentimiento, orientaciones semánticas o polaridades). Ahora estamos listos para definir una opinión.

DEFINICIÓN 3.3. (Opinión) Una opinión (u opinión regular) es un quíntuple, $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$, donde e_i es el nombre de una entidad, a_{ij} es un aspecto de e_i, oo_{ijkl} es la orientación de la opinión sobre el aspecto a_{ij} de entidad e_i, h_k es el titular de la opinión, y t_l es el momento en que la opinión es expresado por h_k. La orientación de la opinión oo_{ijkl} puede ser positiva, negativa o neutral, o expresarse con diferentes niveles de fuerza/intensidad. Cuando una opinión es sobre la entidad misma como un todo, utilizamos el aspecto especial GENERAL para denotarlo.

Objetivo de la minería de opinión: A partir de una colección de documentos de opinión D, descubrir todos los quíntuples de opinión $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ en D. Para lograr este objetivo, es necesario realizar las siguientes tareas:

Tarea 1 (extracción y agrupación de entidades): Extraer todas las expresiones de entidad en D y agrupar las expresiones de entidades sinónimas en grupos de entidades. Cada grupo de expresión de entidad indica una entidad única (e_i).

Tarea 2 (extracción y agrupación de aspectos): Extraer todas las expresiones de aspecto de las entidades y agrupar las expresiones de aspecto en grupos. Cada grupo de expresión de aspecto de la entidad (e_i) indica un aspecto único (a_{ij}).

Tarea 3 (titular de la opinión y extracción de tiempo): extraiga esta información del texto o de los datos no estructurados.

Tarea 4 (clasificación del sentimiento de aspecto): determinar si cada opinión sobre un aspecto es positiva, negativa o neutra.

Tarea 5 (generación quíntuple de opinión): produzca todos los quíntuples de opinión $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ expresados en D en función de los resultados de las tareas anteriores.

Antes de ir más allá, analicemos otros dos conceptos importantes relacionados con la minería de opiniones y el análisis de sentimientos, es decir, la subjetividad y la emoción.

Definición 3.4. (Subjetividad de la oración) Una oración objetiva presenta cierta información factual sobre el mundo, mientras que una oración subjetiva expresa algunos sentimientos, puntos de vista o creencias personales.

Definición 3.5. (Emoción) Las emociones son nuestros sentimientos y pensamientos subjetivos.

1.2. Resumen de Opinión Basada en el Aspecto.

La mayoría de las aplicaciones de minería de opinión deben estudiar las opiniones de un gran número de personas que opinan. Una opinión de una sola persona generalmente no es suficiente para actuar. Esto indica que es deseable algún tipo de resumen de opiniones. Los quintuples de opinión definidos anteriormente proporcionan una excelente fuente de información para generar resúmenes tanto cualitativos como cuantitativos. Una forma común de resumen se basa en aspectos, y se denomina resumen de opinión basado en aspectos (o resumen de opinión basado en características). Para más detalles ver [6].

2. Clasificación del Sentimiento del Documento

Ahora estamos listos para discutir algunos de los principales temas de investigación de la minería de opinión. Esta sección se centra en la clasificación de sentimientos, que ha sido ampliamente estudiada en la literatura. La tarea también se conoce comúnmente como la clasificación de sentimientos a nivel de documento porque considera todo el documento como la unidad de información básica.

Definición 3.6. (Sentimiento de nivel de documento) Dado un documento de opnión d evaluando una entidad e, determinar la orientación de opinión oo en e, es decir, determinar

un aspecto general en el quíntuple $(e, GENERAL, oo, h, t)$. e, h, y t se suponen conocidos o irrelevantes.

Una suposición importante sobre la clasificación de sentimientos es la siguiente: Hipótesis: la clasificación de sentimientos supone que el documento de opinión d (por ejemplo, una revisión de producto) expresa opiniones sobre una sola entidad e y las opiniones son de un único titular de opinión h.

Esta suposición es válida para las revisiones de productos y servicios por parte de los clientes, ya que cada una de esas revisiones generalmente se enfoca en un solo producto y está escrita por un solo revisor. Sin embargo, es posible que no sea válido para publicar en un foro y blog porque en dicha publicación el autor puede expresar opiniones sobre varios productos y compararlos usando oraciones comparativas.

La mayoría de las técnicas existentes para la clasificación de sentimiento a nivel de documento se basan en el aprendizaje supervisado, aunque también hay algunos métodos no supervisados. Revisar [6] para más información.

3. Subjetividad de la Oración y Clasificación de los Sentimientos.

Naturalmente, las mismas técnicas de clasificación de sentimiento a nivel de documento también se pueden aplicar a oraciones individuales. La tarea de clasificar una oración como subjetiva u objetiva a menudo se denomina clasificación de subjetividad en la literatura existente. La resultante oraciones subjetivas también se clasifican como expresiones positivas o negativas opiniones, lo que se denomina clasificación del sentimiento a nivel de la oración.

Definición 3.7. Dada una frase f, se realizan dos subtareas:

(1) Clasificación de subjetividad: determina si f es una oración subjetiva o una oración objetiva,
(2) Clasificación de sentimiento a nivel de oración: si f es subjetiva, determina si expresa una opinión positiva, negativa o neutral.

Tenga en cuenta que el quíntuplo (e, a, oo, h, t) no se utiliza para definir el problema aquí porque la clasificación a nivel de oración es a menudo un paso intermedio. En la mayoría de las aplicaciones, uno necesita saber qué entidades o aspectos de las entidades son los objetivos de las opiniones. Sabiendo que algunas oraciones tienen opiniones positivas o negativas, pero no

sobre qué, tiene un uso limitado. Sin embargo, las dos subtareas siguen siendo útiles porque (1) filtra aquellas frases que no contienen opiniones, y (2) después de saber qué entidades y se habla de aspectos de las entidades en una oración, este paso puede ayudarnos a determinar si las opiniones sobre las entidades y sus aspectos son positivos o negativos.

4. Expansión del Léxico de Opinión.

En las secciones anteriores, mencionamos que las palabras de opinión se emplean en muchas tareas de clasificación de sentimientos. Ahora discutimos cómo se generan esas palabras. En la literatura de investigación, las palabras de opinión también se conocen como palabras de opinión o palabras de sentimiento. Las palabras de opinión positiva se usan para expresar algunos estados deseados, mientras que las palabras de opinión negativas se usan para expresar algunos estados no deseados. Ejemplos de palabras de opinión positivas son: bello, maravilloso, bueno y sorprendente. Ejemplos de las palabras de opinión negativa son malas, pobre y terribles. Además de las palabras individuales, también hay frases de opinión y expresiones idiomáticas, por ejemplo, le cuestan a alguien un brazo y una pierna. Colectivamente, se les llama el léxico de opinión. Son fundamentales para la minería de opiniones por razones obvias.

Para compilar o recopilar la lista de palabras de opinión, se han investigado tres enfoques principales: enfoque manual, enfoque basado en diccionarios y el enfoque basado en corpus. El enfoque manual consume mucho tiempo y, por lo tanto, no se suele usar solo, sino que se combina con enfoques automáticos como control final, porque los métodos automátizados cometen errores. Ver [6].

5. Análisis de Sentimientos Basados en Aspectos.

Aunque la clasificación de textos de opinión a nivel de documento o de frase es útil en muchos casos, no proporciona los detalles necesarios para muchas otras aplicaciones. Un documento con opinión positiva sobre una entidad en particular no significa que el autor tenga opiniones positivas sobre todos los aspectos de la entidad. Del mismo modo, un documento con opiniones negativas no significa que el autor no le gusta todo. En un típico documento de opinión, el autor escribe aspectos positivos y negativos de la entidad, aunque el sentimiento general sobre la entidad puede ser positivo o negativo. La clasificación de documentos y

sentencias no proporciona dicha información. Para obtener estos detalles, tenemos que ir al nivel de aspecto. Es decir, necesitamos el modelo completo de la Sección. 2.1, es decir, minería de opinión basada en aspectos. En lugar de tratar la minería de opiniones simplemente como una clasificación de sentimientos, el análisis de sentimientos basado en aspectos introduce una serie de problemas que requieren capacidades de procesamiento de lenguaje natural más profundas, y también producen un conjunto de resultados más completo. Recuerde que, a nivel de aspecto, el objetivo de minería es descubrir cada quíntuple $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ en un documento dado d. Para lograr el objetivo, se deben realizar cinco tareas. Esta sección se centra principalmente en las siguientes dos tareas principales, que también se han estudiado más extensamente por los investigadores (en [6], se discute brevemente algunas otras tareas):

(1) **Extracción de Aspectos:** Extraer los aspectos que han sido evaluados. Por ejemplo, en la oración, "La calidad de imagen de esta cámara es asombrosa", el aspecto es "calidad de imagen" de la entidad representada por "esta cámara". Tenga en cuenta que "esta cámara" no indica el aspecto GENERAL porque la evaluación no se trata de la cámara como un todo, sino de su calidad de imagen. Sin embargo, la frase "Me encanta esta cámara" evalúa la cámara como un todo, es decir, el aspecto GENERAL de la entidad representada por "esta cámara". Tenga en cuenta siempre que hablemos de un aspecto, debemos saber a qué entidad pertenece. En nuestra discusión a continuación, a menudo omitimos la entidad solo por simplicidad de presentación.

(2) **Clasificación del Sentimiento de Aspecto:** determinar si las opiniones sobre diferentes los aspectos son positivas, negativas o neutrales. En el primer ejemplo anterior, la opinión sobre el aspecto de "calidad de imagen" es positiva, y en el segundo ejemplo, la opinión sobre el aspecto GENERAL también es positiva.

6. Minería Opiniones comparativas.

Expresar directa o indirectamente opiniones positivas o negativas sobre una entidad y sus aspectos es solo una forma de evaluación. Comparar la entidad con algunas otras entidades similares es otra. Las comparaciones están relacionadas, pero también son bastante diferentes de las opiniones regulares. No sólo tienen diferentes significados semánticos, sino también diferentes formas sintácticas. Por ejemplo, una frase de opinión típica es "La calidad de

imagen de esta cámara es excelente", y una oración comparativa típica es "La calidad de imagen de Camera-x es mejor que la de Camera-y".

Tipos de relaciones comparativas: las relaciones comparativas o las comparaciones se pueden agrupar en cuatro tipos principales. Los primeros tres tipos se llaman comparaciones calificables y el último las comparaciones no calificables.

(1) *Comparaciones calificables no iguales:* las relaciones del tipo mayor o menor que expresan un orden de algunas entidades con respecto a algunos de sus aspectos compartidos, por ejemplo, *"El chip Intel es más rápido que el de AMD".* Este tipo también incluye las preferencias del usuario, por ejemplo, *"Prefiero Intel a AMD".*

(2) *Comparaciones equitativas:* Relaciones del tipo igual a ese estado, dos o más entidades son iguales con respecto a algunos de sus aspectos compartidos, por ejemplo, *"El rendimiento de Carro-x es aproximadamente el mismo que el de Carro-y".*

(3) *Comparaciones superlativas:* Relaciones del tipo mayor o menor que todas las demás que clasifican una entidad sobre todas las demás, por ejemplo, *"El chip Intel es el más rápido".*

(4) *Comparaciones no calificables:* Relaciones que comparan aspectos de dos o más entidades, pero no las califican. Hay tres subtipos principales:

- La entidad A es similar o diferente de la entidad B con respecto a algunos de sus aspectos comunes, por ejemplo, *"El sabor de la Coca-Cola es diferente al de la Pepsi".*
- La entidad A tiene el aspecto a_1, y la entidad B tiene el aspecto a_2 (a_1 y a_2 son generalmente sustituibles), por ejemplo, *"las PC de escritorio usan parlantes externos pero las laptops usan parlantes internos".*
- La entidad A tiene aspecto a, pero la entidad B no tiene, por ejemplo, *"Teléfono-x tiene un auricular, pero Teléfono-y no tiene".*

Las palabras comparativas utilizadas en comparaciones de calificaciones no iguales se pueden categorizar en dos grupos según expresen cantidades aumentadas o disminuidas, que son útiles en el análisis de opinión.

- *Comparativas crecientes:* Tal comparación expresa una mayor cantidad, por ejemplo, más y más larga.

- *Comparativas decrecientes:* Tal comparación expresa una cantidad disminuida, por ejemplo, menos y muy poco.

Objetivo de la minería de opiniones comparativas: Dada una colección de documentos de opinión D, descubra en D todos los sextuples de opinión comparativa de la forma (E_1, E_2, A, EP, h, t), donde E_1 y E_2 son los conjuntos de entidades que se comparan en función de sus aspectos compartidos A (las entidades en E_1 aparecen antes que las entidades en E_2 en la oración), $EP(\in \{E_1, E_2\})$ es el conjunto de entidades preferidas del ponente de la opinión h, y t es el momento en que se expresa la opinión comparativa.

Ejemplo 3.8. Considere la frase comparativa "La óptica de Canon es mejor que las de Sony y Nikon", escrita por John en 2010. La opinión comparativa extraída es: ({Canon}, {Sony, Nikon}, {óptica}, preferida: {Canon}, John, 2010) La entidad que establece E_1 es {Canon}, el conjunto de entidades E_2 es {Sony, Nikon}, su conjunto de aspectos compartidos A que se compara {óptica}, el conjunto de entidades preferido es {Canon}, el portador de la opinión h es John y el momento t cuando se escribió esta opinión comparativa es 2010.

7. Detección de Spam de Opinión.

Se ha convertido en una práctica común para las personas encontrar y leer opiniones en la Web para muchos propósitos. Por ejemplo, si uno quiere comprar un producto, normalmente va a un comerciante o a un sitio de reseñas (por ejemplo, amazon.com) para leer algunas reseñas de usuarios existentes del producto. Si uno ve muchas críticas positivas del producto, es muy probable que compre el producto. Sin embargo, si uno ve muchas críticas negativas, lo más probable es que elija otro producto. Las opiniones positivas pueden generar ganancias y/o famas financieras significativas para organizaciones e individuos. Desafortunadamente, esto ofrece buenos incentivos para el spam de opinión, que se refiere a actividades humanas (por ejemplo, comentarios sobre spam) que intentan engañar deliberadamente a los lectores o sistemas automatizados de minería de opinión emitiendo opiniones positivas inmerecidas a algunas entidades objetivo, con el fin de promover las entidades y/o dando opiniones negativas injustas o falsas a otras entidades para dañar su reputación. Dichas opiniones también se llaman opiniones falsas, opiniones ficticias o críticas falsas.

Spammers individuales y spammers grupales: un spammer puede actuar individualmente (por ejemplo, el autor de un libro) o como miembro de un grupo (por ejemplo, un grupo de empleados de una empresa).

7.1. Detección de spam basada en aprendizaje supervisado.

En general, la detección de spam puede formularse como un problema de clasificación con dos clases, spam y no spam. Sin embargo, etiquetar manualmente los datos de entrenamiento para el aprendizaje es muy difícil, si no imposible. El problema es que identificar los comentarios de spam simplemente leyendo los comentarios es extremadamente difícil porque un spammer puede elaborar cuidadosamente un comentario de spam que es como cualquier otro comentario inocente.

La regresión logística se utilizó para la construcción de modelos. Los resultados experimentales mostraron algunos resultados interesantes. Ver [6].

7.2. Detección de spam basada en comportamientos anormales.

Debido a la dificultad de etiquetar manualmente los datos de capacitación, el tratamiento de la detección de spam de opinión como un problema de aprendizaje supervisado es problemático porque muchas revisiones no duplicadas también pueden ser spam. Aquí, describimos dos técnicas que intentan identificar comportamientos atípicos de los revisores para detectar spammers. Por ejemplo, si un revisor escribió todas las revisiones negativas para una marca, pero otros revisores fueron todos positivos sobre la marca, entonces este revisor es, naturalmente, un sospechoso de spam.

7.3. Detección de Spam Grupal.

Un algoritmo de detección de spam grupal encuentra grupos de spammers que trabajan juntos para promover o degradar algunos productos.

Capítulo 4

Resultados y Análisis.

En esta sección, haremos un análisis de minería de opinión respecto a la pobreza multidimensional en Venezuela con información extraída de la red social Twitter, de la misma manera haremos un análisis de minería de opinión para ingresos percibidos. Cabe destacar que estaremos enfocado en las opiniones de los usuarios y no en un índice de pobreza multidimensional per se.

Iniciaremos con el análisis de sentimientos, luego aplicaremos LDA, identificaremos los actores de los tweets, y finalmente, la ubicación geográfica de emisión de los titulares de los tweets.

1. Fuentes de los Datos.

Los datos provienen de la extracción de tweets mediante la API de Twitter (ver detalles en el Anexo III), con un tiempo comprendido desde el 30 de noviembre de 2016 al 20 de septiembre de 2018. Hemos obtenido la cantidad de 7670 tweets. La clave de búsqueda utilizada en la API de Twitter fue, "(pobreza OR pobres OR pobreza extrema) lang:es place_country:VE."

En el Cuadro 1. Observamos la cantidad de tweets por cada dimensión de pobreza multidimensional e ingresos.

Los datos del Cuadro 1. Fueron obtenidos mediante un diccionario[1] de palabras que permite clasificar las dimensiones. Del Cuadro 1. Podemos observar que la dimensión salud de pobreza arrojó un 19,17 %, mostrando en particular que los usuarios se inclinan hacia la opinión referente a salud, en las secciones posteriores analizaremos el sentido de la opinión no solo de la dimensión salud, si no para cada una de las clasificaciones tales como, dimensión educación, dimensión condiciones de vida e ingresos. La idea es evaluar la pobreza multidimensional y los ingresos percibidos, tomando en cuenta que la pobreza no solo debe enfocarse

[1] ver Capítulo 1. Pág. 8

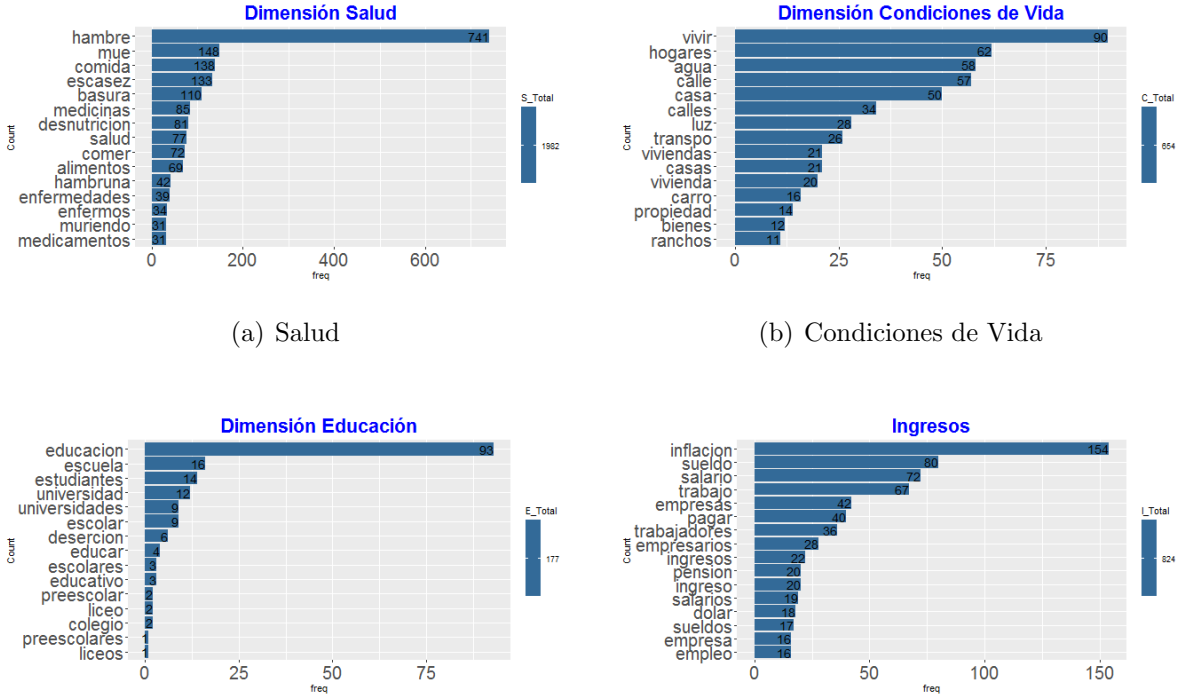
Descripción	Cantidad de tweets	Porcentaje
Total	7670	100
Total por Clasificación	2903	37,85
Total por Dimensiones	2212	28,83
Dimensión Salud	1471	19,17
Ingresos	691	9,0
Dimensión Condiciones de Vida	578	7,53
Dimensión Educación	163	2,12

CUADRO 1. Cantidad de tweets por cada clasificación.

en el ingreso percibido, sino en las carencias presenciadas en la vida de las personas. Por ejemplo, una persona pobre puede tener una mala salud o estar desnutrida, puede carecer de agua limpia o electricidad, tener un trabajo precario o tener muy poca educación. Enfocarse en un solo factor, tal como el ingreso, no es suficiente para capturar la verdadera realidad de la pobreza.

Antes de continuar veamos el siguiente gráfico que contiene las frecuencias de palabras por clasificación de dimensión e ingresos.

En la Figura 4.1. Se aprecia en temas de salud en la subcategoría de nutrición las palabras hambre (741), comida (138), desnutrición (81), comer (72), alimentos (69), hambruna (42), para un total de 1.143 coincidencias de palabras referentes solo al hambre. Por otro lado, la dimensión condiciones de vida representa al menos una frecuencia de 185 palabras referente a Vivienda, distribuidas de la siguiente manera: hogares (62), Casas (71), viviendas (41) y ranchos (11). Seguidamente las palabras mas frecuentes con respecto a la dimensión educación relacionan a la institución y al estudiante. Finalmente, en la categoría ingresos las palabras con más coincidencias la hemos contabilizado como sigue: sueldos (97), salarios (91), ingresos (20), para un total de 208 palabras mas frecuentes con opinión en salarios, 154 repeticiones de palabras en temas de inflación y 52 repeticiones de palabras respecto al trabajo.



(a) Salud

(b) Condiciones de Vida

(c) Educación

(d) Ingresos

FIGURA 4.1. Frecuencias de palabras.

2. Análisis de Sentimiento.

Para el análisis de sentimiento se hizó uso del un léxico Afinn (Un diccionario). Las palabras que son percibidas de manera negativa tinen puntuaciones de -4 a -1; y las positivas de 1 a 4. La versión utilizada es una traducción automática, de inglés a español, de la versión del léxico presente en el conjunto de datos sentiments de tidytext, paquete de R, con algunas correcciones manuales, esto quiere decir que este léxico tendrá algunos defectos, pero será suficiente para nuestro análisis.

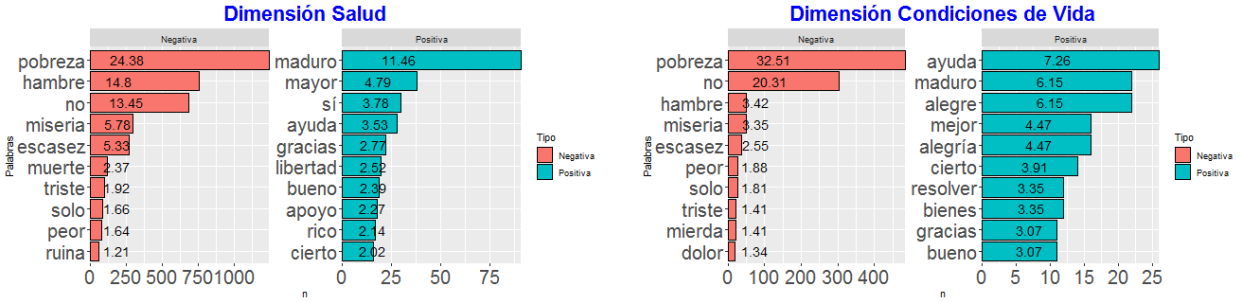
En este proceso vamos a contestar las siquientes preguntas:

¿Cuáles palabras han influido para determinar los sentimientos?, ¿Qué sentimientos han sido predominantes? ¿Positivo, negativo? y ¿Cómo han cambiado los sentiminetos a través del tiempo?.

2.1. ¿Cuáles palabras han influido para determinar los sentimientos?

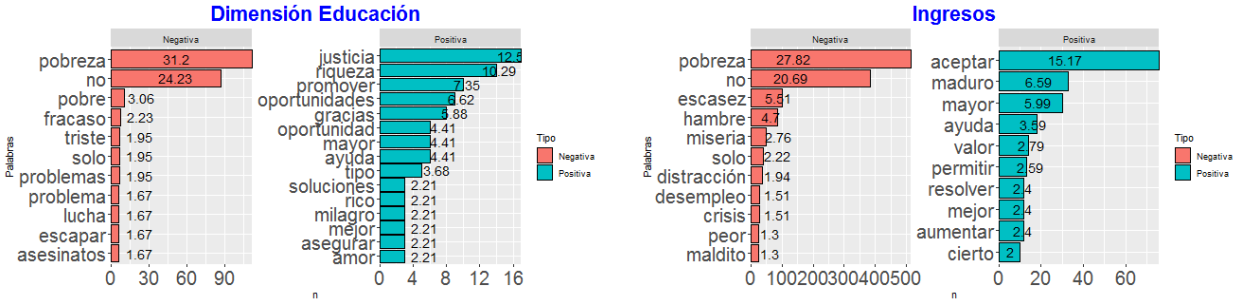
Para responder esta pregunta presentaremos cuatros gráficas donde se aprecia las palabras influenciadas por cada dimensión de pobreza e ingresos, palabras que definen el sentimiento

general de la opinión. Se presenta la gráfica con las 10 primeras palabras más frecuentadas por cada clasificación.



(a) Salud

(b) Condiciones de Vida



(c) Educación

(d) Ingresos

FIGURA 4.2. Palabras que influencian los sentimientos. Los pesos están dados por la frecuencia de cada palabra y el total de la suma de frecuencias de las palabras positivas y negativas.

La Figura 4.2. Muestra las palabras más influyentes en sentimientos negativo tales como: hambre, miseria, escasez, triste, peor, problemas, muerte, desempleo; en sentimientos positivos las palabras más influyentes son: ayuda, libertad, apoyo, alegre, resolver, bienes, justicia, promover, oportunidades, soluciones, permitir, resolver, aumentar.

Mediante los resultados del gráfico 4.2. Notamos que la dimensión salud pone en manifiesto que las personas no se están alimentando bien y por ende están propensos a presentar desnutrición, por su parte en las palabras positivas hay nociones de ayuda y apoyo lo cual refleja la posibilidad de un problema de salud. En dimensión condiciones de vida se aprecia tristeza, miseria y escasez y su contra parte alegría, resolución y bienes, lo que pone en manifiesto la carencia, posiblemente por falta de bienes necesarios para tener una calidad

de vida. La dimensión educación hace mención de problemas con la educación y necesidad de generar oportunidades de estudio. Por último, en ingresos se aprecia la posibilidad del desempleo y la intención de una forma de solucionarlo.

2.2. ¿Qué sentimientos han sido predominantes? ¿Positivo, negativo?

Esta pregunta la contestaremos con el siguiente gráfico. Note que las palabras con sentimientos negativos predominan en todas las clasificaciones. La dimensión condiciones de vida

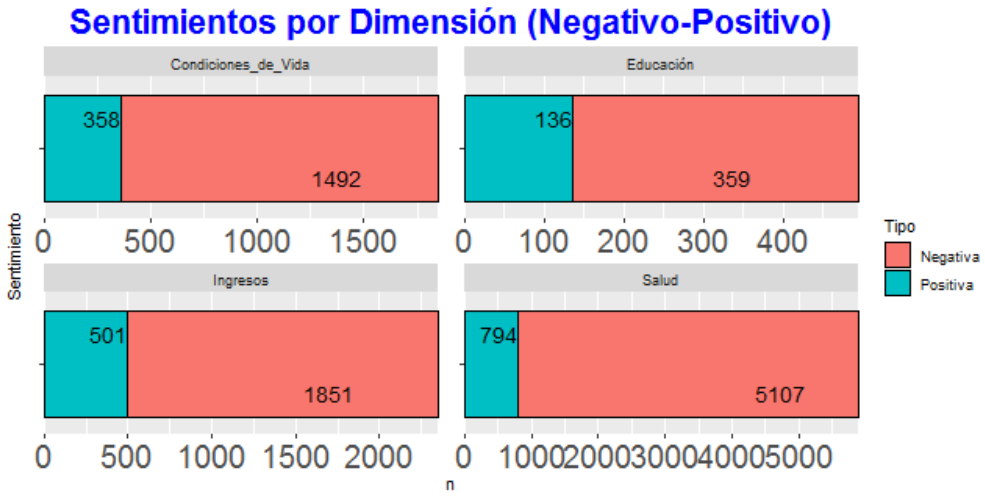


FIGURA 4.3. Sentimientos predominantes Positivo - Negativo.

contabilizó al menos 358 palabras positivas y 1492 palabras negativas, dimensión educación arrojó 136 palabras positivas y 359 palabras negativas, ingresos registró 501 palabras positivas y 1851 palabras negativas, dimensión salud presentó 794 y 5107 palabras positiva y negativa respectivamente.

2.3. ¿Cómo han cambiado los sentiminetos a través del tiempo?

Con el fin de evaluar como varían las opiniones a través del tiempo se presentan dos gráficas. La primera (Figura 4.4.) media de las puntuaciones (entre -4 y 4) de las opiniones negativo - positivo y la segunda (Figura 4.5.) una regresión local para la media de las puntuaciones por días. El resultado es que los sentimientos presentan tendencias negativas en el tiempo registrado, 30 de niviembre de 2016 al 20 de agosto de 2018.

También se realizó una regresión lineal y una regresión local para las puntuaciones, arrojando los mismos resultados de tendencia negativa.

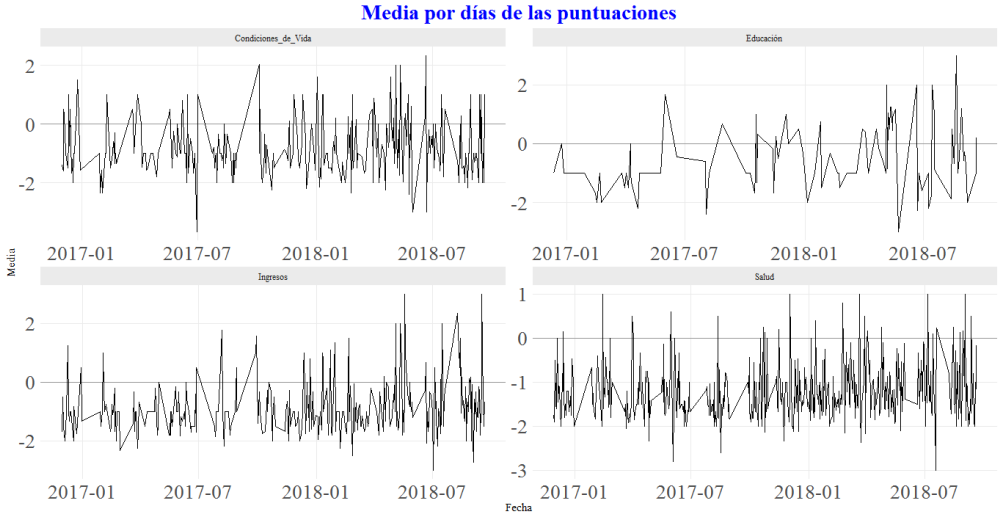


FIGURA 4.4. Puntuaciones (sentimientos predominantes Positivo - Negativo).

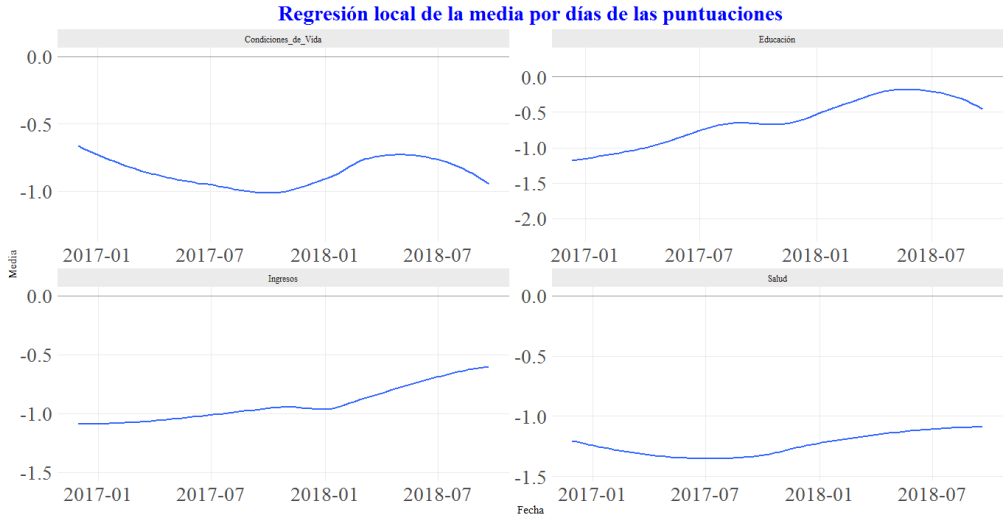


FIGURA 4.5. Regresión local de la media por días de las puntuaciones (sentimientos predominantes Positivo - Negativo).

Nota: En la regresión lineal local se hizo ajuste locales de polinomios.

3. Análisis de Tópicos.

En esta sección analizaremos los tópicos latentes de Dirichlet para cada dimensión de pobreza e ingresos. Tendremos la finalidad identificar los siguientes puntos:

(1) Aspectos generales del tópico: relaciones con otros temas o actores claves, causalidad, efectos, Instituciones o actores involucrados, medios de superación, entre otros.
(2) Identificación de demandas, necesidades o problemas que los usuarios de la red relacionan con pobreza multidimensional.

En el cuadro 2, se presentan los tópicos latentes, 5 tópicos de 10 palabras, para la dimensión salud. Veamos las palabras que definen los tópicos, por ejemplo, tópico 1 las palabras relacionadas son; hambre, miseria, pobreza, hambruna, maduro, nada. En el tópico 2 tenemos palabras como; pobreza, extrema, venezolanos, basura, comida, medicinas, niños, alimentos. El tópico 3 tenemos como palabras más influyentes; pobreza, pueblo, escasez, gobierno, hambre, corrupción, desnutrición, inflación, falta. En el tópico 4 tenemos las palabras; pobres, hambre, comer, menos, comida, contra. Tópico 5 refleja la relación entre las palabras; pobreza, Venezuela, régimen, mientras, miseria, Chávez, vida.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
hambre	pobreza	pobreza	pobres	pobreza
mas	extrema	mue	hambre	venezuela
miseria	venezolanos	pueblo	angos	pais
pobreza	basura	escasez	hoy	hay
maduro	comida	gobierno	comer	regimen
dia	medicinas	hambre	ellos	mientras
cada	gente	corrupcion	menos	hasta
ahora	ningos	desnutricion	comida	miseria
hambruna	millones	inflacion	pobre	chavez
nada	alimentos	falta	contra	vida

CUADRO 2. Tópicos de discusión dimensión salud.
Nota: Por motivos de códigos en R hemos cambiado ñ por ng y hemos eliminados los acentos.

Los tópicos del Cuadro 2., dejan al descubierto la escasez de alimentos, por lo cual existe la posibilidad de una alimentación pírrica, la causa de este problema es la corrupción y la inflación, los actores mencionados son: Maduro y Chávez, y los sujetos como niños, venezolanos, gobierno y régimen, definen el sentido de la opinión.

En la Gráfica 4.6., se aprecia que el tópico 1 y 4 son los más influyente a través del tiempo estudiado, los tópicos 2, 3 y 5 se mantiene poco variante en el tiempo.

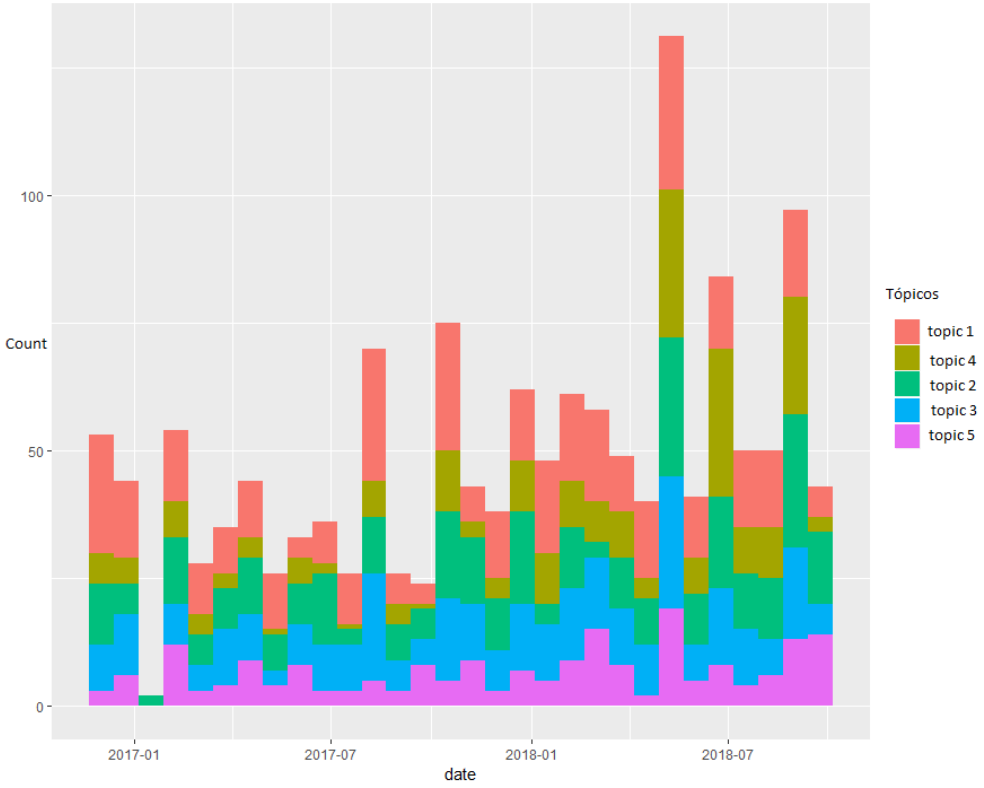


FIGURA 4.6. Cambio de los tópicos en función del tiempo (dimensión salud).

A continuación, veamos los tópicos de la dimensión Condiciones de Vida. El cuadro 3, muestra cuatros tópicos de 7 palabras cada uno, que reflejan la conexión entre las palabras pobreza, hogares, gobierno, agua, luz, venezolanos, calles, niños, transporte y casa. Estos tópicos reflejan la falta de servicios de agua, luz y transporte. También, se menciona que hay personas en la calle (en busca de comida o por falta de un hogar, no queda claro), pero están en las calles, se identifica sujetos como: niños, gobierno y venezolanos.

Topic 1	Topic 2	Topic 3	Topic 4
pobreza	pobres	pobreza	mas
pais	hay	vivir	venezuela
hogares	agua	hambre	casa
extrema	millones	miseria	calle
pueblo	comida	venezolanos	ningos
estan	luz	gente	transpo
gobierno	etc	calles	cada

CUADRO 3. Tópicos de discusión dimensión condiciones de vida.
Nota: Por motivos de códigos en R hemos cambiado ñ por ng y hemos eliminados los acentos.

En la Gráfica 4.7., podemos notar que los tópicos 4 y 2 son los más resaltantes en el tiempo. Por otra parte, los tópicos 1 y 3 se observa con poca variación a través del tiempo estudiado.

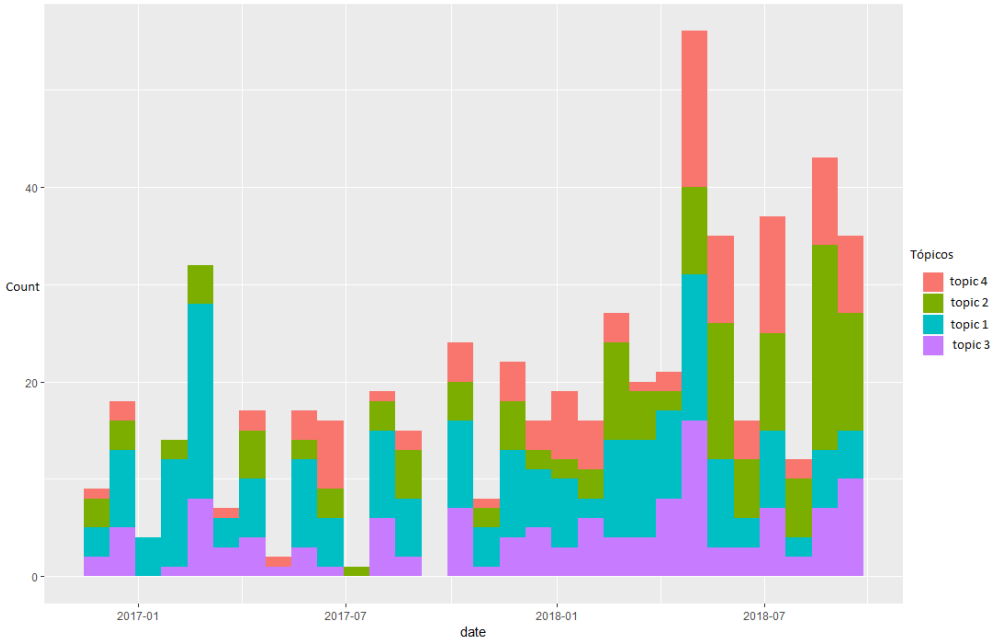


FIGURA 4.7. Cambio de los tópicos en función del tiempo (dimensión condiciones de vida).

Ahora veamos los tópicos de la dimensión Educación. Las palabras pobreza, Venezuela, país, estudiantes, justicia, salud, derecho, educación, calidad, universidad, equitativa y promover; están involucradas en los cuatros tópicos de dimensión educación. Para más detalles véase el cuadro 4. Entre los tópicos tenemos las opiniones como: "Los estudiantes son la riqueza de un país", "derecho a una escuela", "educación de calidad" y "promover oportunidad equitativa de estudio". Estas opiniones nos dejan en claro que hay un problema con la educación, si bien no es evidente identificar el problema, logramos encontrar en los tweets afirmaciones como: "deserción estudiantil", "malas condiciones de algunas instituciones educativas" y "falta de educadores calificados para dictar clases". Como consecuencia, si se mantienen estas condiciones, tendremos una escolaridad muy baja en un futuro no muy lejano. No hemos identificado actores relevantes en los tópicos.

Topic 1	Topic 2	Topic 3	Topic 4
pobreza	pobres	buena	educacion
venezuela	mas	escuela	calidad
paises	hoy	hay	universidad
riqueza	justicia	nuevo	opo
pais	mundo	salud	durante
salir	moral	derecho	equitativa
estudiantes	millones	pobre	promover

CUADRO 4. Tópicos de discusión dimensión educación.
Nota: Por motivos de códigos en R hemos cambiado ñ por ng y hemos eliminados los acentos.

La Figura 4.8., muestra que los tópicos no son muy influyentes en el tiempo de estudio. Sin embargo, los usuarios están siempre hablando del tema de educación, lo cual no deja de ser importante cuando de educación se trata, considerando que en ocasiones los jóvenes no ven con buenos ojos la intención de estudiar. También de la Gráfica 4.8., se observa a partir de enero del 2018 una participación casi equitativa de los cuatro tópicos presentados. Este resultado infiere que los usuarios prefieren estar bien de salud y poseer bienes antes que estudiar.

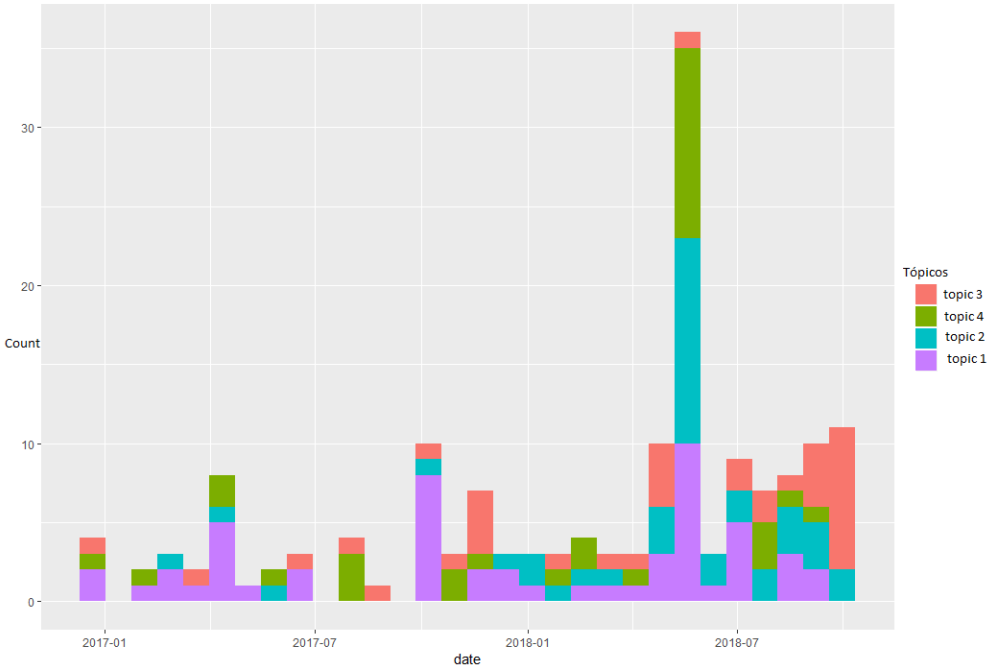


FIGURA 4.8. Cambio de los tópicos en función del tiempo (dimensión educación).

Finalmente, en el Cuadro 5, presentamos los tópicos de discusión referente a los ingresos percibidos, las palabras más influyentes son aumento, trabajadores, pagar, migajas, inflación y debe. Claramente deja evidenciado que los salarios percibidos son bajos.

Topic 1	Topic 2	Topic 3	Topic 4
inflacion	sueldo	mas	pobreza
pobreza	salario	pobres	miseria
hambre	podemos	pueblo	trabajo
venezuela	regimen	trabajadores	hay
pais	extrema	aumento	menos
empresas	aumentos	ahora	cada
escasez	pagar	empresarios	estos
economica	migajas	ricos	debe

CUADRO 5. Tópicos de discusión ingresos.

Nota: Por motivos de códigos en R hemos cambiado ñ por ng y hemos eliminados los acentos.

Los temas más relevantes detectados en ingresos percibidos son: salarios bajos y poca posibilidad de empleo, producto de una mala economía, la inflación y escasez de vacantes en las empresas. No identificamos actores en los tópicos de discusión.

La Gráfica 5.9., refleja que los tópicos más influyentes son 1 y 3, y los pocos influyentes, pero presentes en casi todo el tiempo estudiado son los tópicos 4 y 2.

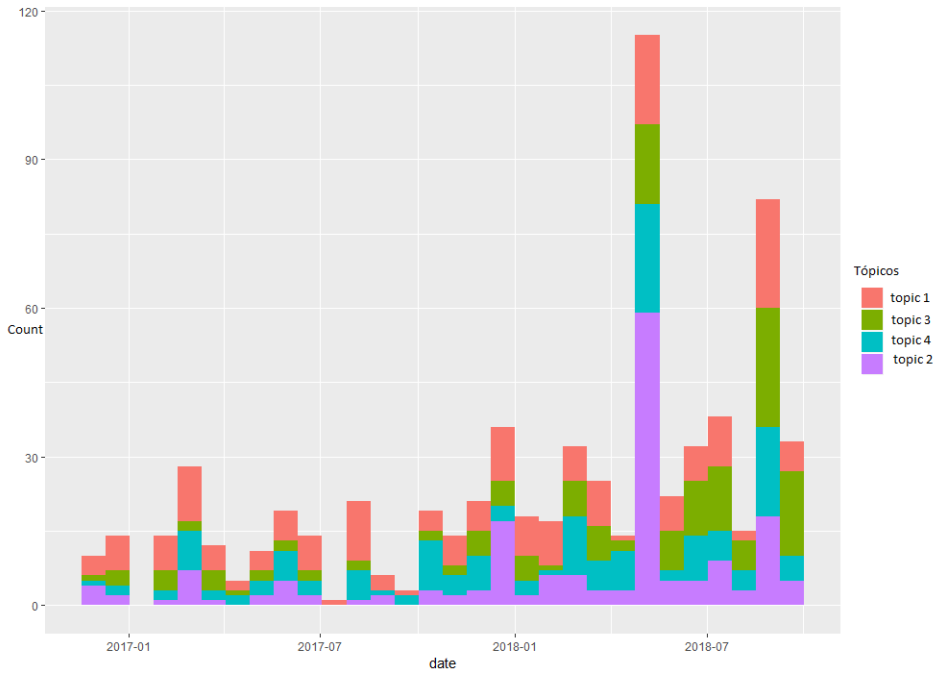


FIGURA 4.9. Cambio de los tópicos en función del tiempo (ingresos).

3.1. ¿Retweet o únicos.?

Dado que el texto de un tweet puede ser Retweet o ser un único tweet, hemos desarrollado un algoritmo que permite identificar por cada tópico, y por cada dimensión, si se está en presencia de un Retweet o no. Como resultado obtuvimos que, de cada 100 tweet, 3 son Retweet. Esta observación se repite por cada tópico presentado anteriormente en la Sección 3 de este Capítulo. Puesto que la mayoría de los tweets son únicos no tenemos una información replicada por parte de los usuarios, lo cual resulta útil ya que la opinión es personalizada. Aunque vale la pena decir que si se trata de un Retweet el usuario está apoyando o está de acuerdo con cierta idea.

3.2. Emisor del tweet.

Las opiniones positivas pueden generar ganancias y/o famas financieras significativas para organizaciones e individuos. Desafortunadamente, esto da buenos incentivos para el spam de opinión, que se refiere a actividades humanas (por ejemplo, comentarios sobre empresas o productos) que intentan engañar deliberadamente a los lectores o sistemas automatizados de minería de opinión emitiendo opiniones positivas inmerecidas a algunas entidades objetivo, con el fin de promover las entidades y/o dando opiniones negativas injustas o falsas a otras entidades para dañar su reputación. Dichas opiniones también se llaman opiniones falsas, opiniones ficticias o críticas falsas. Ver [6].

En virtud de las opiniones Spam, hemos observado los titulares de opinión contenidos en nuestra muestra, la observación se ha realizado con la ayuda de la herramienta Python mirando uno a uno el titular que emite el tweet. En esta realización encontramos una variedad de usuarios distribuidos en: medios de comunicación, personas naturales, personas que manifiestan ser afectos al gobierno, instituciones del estado, empresas privadas entre otros. Dado que estamos ubicados en Venezuela es de interés contabilizar las personas afectas al gobierno y las instituciones del estado que están interactuando con el tema de pobreza multidimensional. En el Cuadro 6, se puede apreciar que solo 208 de 2051 usuarios, es decir, 10,14 %, se identifican con mucho atino al actual gobierno venezolano. La distribución consta de 166 (8,09 %) usuarios afectos al gobierno y 42 (2,05 %) instituciones del estado.

Descripción		**Persona afecta al Gobierno**		**Institución del Estado**		**Total**	
	Total de usuarios	Cantidad de usuarios	%	Cantidad de usuarios	%	Cantidad de usuarios	%
Salud	952	63	6,62	16	1,68	79	8,30
Ingresos	516	44	8,53	5	0.97	49	9.50
Condiciones de Vida	450	41	9,11	15	3,33	56	12,44
Educación	133	18	13,53	6	4,51	24	18,05
Total	2051	166	8,09	42	2,05	208	10,14

CUADRO 6. Cantidad neta de Usuarios que emiten tweet por cada clasificación

3.3. Geolocalización del tweet.

La proyección de la población en Venezuela para el año 2018 según el Instituto Nacional de Estadística (INE) es un poco más de 39 mil millones de habitantes, al menos 32 mil millones que representa el 80 % de la población total se encuentra distribuidos en los estados Zulia, Miranda, Carabobo, Lara, Aragua, Bolívar, Anzoátegui, Táchira, Sucre, Falcón, Portuguesa y el Distrito Capital. Estos 32 mil millones de habitantes ocupa una superficie 447.147 km^2, es decir, 48,7 % del total de la superficie cuadrática de Venezuela, con un densidad de población a lo sumo de 90 %.

En nuestro caso estamos interesados en ubicar geográficamente de donde provienen las opiniones de los usuarios. En la Gráfica 4.10. Se aprecia que las opiniones son emitidas en las regiones donde hay más cantidad de habitantes, siendo este resultado lógico dada la distribución de la población. Los usuarios manifiestan carencia en las dimensiones de pobreza y en ingresos percibidos en las áreas más pobladas de Venezuela, que si bien no es representativo comparar la cantidad de usuarios contenidos en nuestra muestra con el total de la población, nos indica que de alguna manera los usuarios presencian ciertas necesidades asociadas con las dimensiones de pobreza.

En La Figura 4.11. Están geolocalizado 482 tweets asociado a un tópico de salud que contiene las palabras hambre, pobreza, país, Maduro, años, gobierno y miseria.

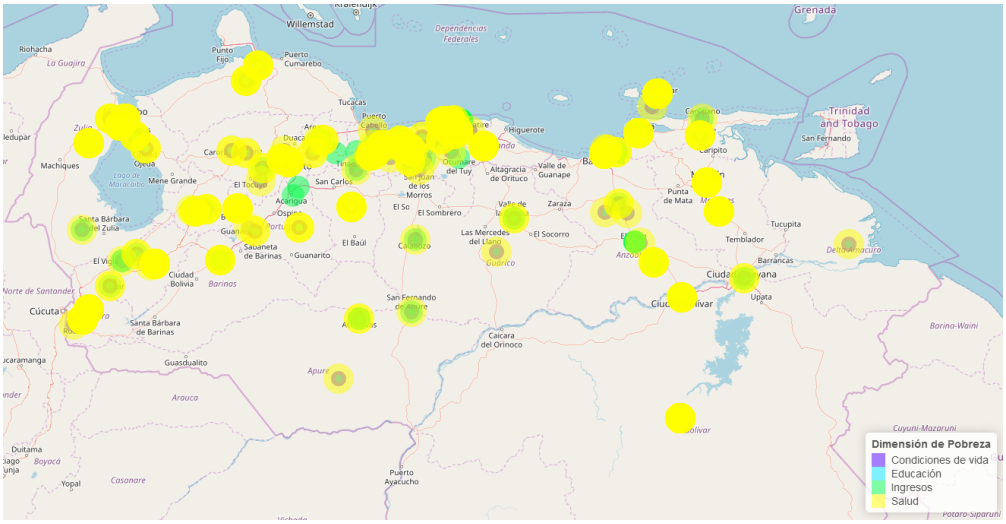


FIGURA 4.10. Geolocalización por dimensión de pobreza e ingresos.

FIGURA 4.11. Geolocalización para un tema de opinión de salud.

La Figura 4.11. Indica en el circulo cuantos tweets son emitidos desde la geolocalización del tweet. Note que existe una asociación de hambre, miseria y pobreza con el gobierno venezolano, tomando en cuenta que la palabra país hace referencia a Venezuela, ya que la captura de los tweets son tomados desde Venezuela.

Es importante resaltar que no se geolocalizaron todos los tweets, motivado a que no fue posible obtener la información aportada por la API de Twitter (puede que el usuario no active la opción de geolocalización de su equipo). En nuestro caso para cada dimensión de pobreza solo logramos geolocalizar al menos el 50 % de los tweets. En general fue posible geolocalizar al menos el 50 % de la muestra.

Conclusión

En general las opiniones poseen tendencias negativas en la medida que transcurre el tiempo. Se identificó que en las opiniones hay más palabras negativas que palabras positivas, en promedio el 20 % del total de palabras identificadas con el diccionario afín son positivas y 79 % son negativas.

Los usuarios manifiestan la existencia de necesidad en temas de salud. Por ejemplo, se mencionó en muchas ocasiones la palabra "hambre". Los usuarios dan más relevancia a los temas de salud e ingresos que a temas asociados a la educación. Se aprecia en varios tópicos de discusión la asociación de pobreza con el gobierno venezolano.

El conjunto de datos estudiado no posee muchos retweets, de cada 100 tweets solo 3 son retweets. Notamos poca influencia de instituciones públicas o personas afectas al gobierno.

Como era de esperar la mayoría de los tweets están ubicados en las áreas más pobladas de Venezuela, evidenciando la percepción de pobreza multidimensional en Venezuela por parte de los usuarios. Por lo menos el 50 % de los tweets se logró geolocalizar.

Como proyecto a futuro deseamos mejorar y ampliar el diccionario de palabras para clasificar las sentimientos positivos y negativos, mediante una red neuronal. Automatizar la identificación de las dimensiones de pobreza y la clasificación de los usuarios que interactúan. Implementar el uso del ecosistema Hadoop para mejorar la capacidad de almacenamiento y el manejo de los datos, así como la visualización de estos de una manera más eficiente.

Anexos

1. Anexo I: Herramienta Python

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible.

Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, usa tipado dinámico y es multiplataforma.

Es administrado por la Python Software Foundation. Posee una licencia de código abierto, denominada Python Software Foundation License, que es compatible con la Licencia pública general de GNU a partir de la versión 2.1.1, e incompatible en ciertas versiones anteriores.

Python fue creado a finales de los ochenta por Guido van Rossum en el Centro para las Matemáticas y la Informática (CWI, Centrum Wiskunde Informatica), en los Países Bajos, como un sucesor del lenguaje de programación ABC, capaz de manejar excepciones e interactuar con el sistema operativo Amoeba.

Unas de las librerías empleadas en este proyecto se pueden apreciar en el cuadro que se presenta a continuación.

Paquete	Descripción
NumPy	Agrega mayor soporte para vectores y matrices, constituyendo una biblioteca de funciones matemáticas de alto nivel para operar con esos vectores o matrices.
pandas	Datos tabulares con columnas de tipo heterogéneo, como en una tabla SQL o en una hoja de cálculo de Excel. Datos matriciales arbitrarios (homogéneamente tipados o heterogéneos) con etiquetas de fila y columna. Cualquier otra forma de conjuntos de datos observacionales / estadísticos. Los datos en realidad no necesitan ser etiquetados para ser colocados en una estructura de datos de pandas.

Cuadro 7. Librerías de Python.

2. Anexo II: Herramienta R

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico.

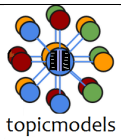
R es una implementación de software libre del lenguaje S pero con soporte de alcance estático. Se trata de uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo y gráficas.

R es parte del sistema GNU y se distribuye bajo la licencia GNU GPL. Está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.

Fue desarrollado inicialmente por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en 1993. Sin embargo, si se remonta a sus bases iniciales, puede decirse que inició en los Bell Laboratories de ATT y ahora Alcatel-Lucent en Nueva Jersey con el lenguaje S. Este último, un sistema para el análisis de datos desarrollado por John Chambers, Rick Becker, y colaboradores diferentes desde finales de 1970. La historia desde este punto es prácticamente la del lenguaje S.

Al igual que S, se trata de un lenguaje de programación, lo que permite que los usuarios lo extiendan definiendo sus propias funciones. De hecho, gran parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionalmente exigentes es posible desarrollar bibliotecas en C, C++ o Fortran que se cargan dinámicamente. Los usuarios más avanzados pueden también manipular los objetos de R directamente desde código desarrollado en C. R también puede extenderse a través de paquetes desarrollados por su comunidad de usuarios.

Algunas de las librerías empleadas en este proyecto son las siguientes:

Paquete	Descripción
	Permiten manipular caracteres individuales dentro de las cadenas en vectores de caracteres, eliminar y manipular espacios en blanco, reconocen cuatro motores de descripción de patrón. El más común es expresiones regulares.
	Ofrece un potente lenguaje de gráficos para crear tramas elegantes y complejas. Permite crear gráficos que representan datos numéricos y categóricos tanto univariados como multivariantes de una manera directa. El agrupamiento se puede representar por color, símbolo, tamaño y transparencia.
	Análisis rápido y fácil uso de datos de fecha y hora, extracción y actualización de componentes de una fecha y hora (años, meses, días, horas, minutos y segundos), manipulación algebraica en objetos de fecha y hora y de lapso de tiempo.
	Es un sistema coherente de paquetes para la manipulación, exploración y visualización de datos que comparten una filosofía de diseño común.
	Interfaz para aplicar funciones de transformación (también denotadas como asignaciones) a corpus. Construye o coacciona una matriz de documento de término o una matriz de término de documento.
	Estima un modelo LDA utilizando, por ejemplo, el algoritmo de VEM o el muestreador de Gibbs.

CUADRO 8. Librerías de R.

3. Anexo III: Herramienta Twitter

Twitter es un servicio de microblogging, con sede en San Francisco, California, con filiales en San Antonio (Texas) y Boston (Massachusetts) en Estados Unidos. Twitter, Inc. fue creado originalmente en California, pero está bajo la jurisdicción de Delaware desde 2007. Desde que

Jack Dorsey lo creó en marzo de 2006, y lo lanzó en julio del mismo año, la red ha ganado popularidad mundial y se estima que tiene más de 500 millones de usuarios, generando 65 millones de tuits al día y maneja más de 800.000 peticiones de búsqueda diarias. Ha sido denominado como el ≪ SMS de Internet ≫.

La red permite enviar mensajes de texto plano de corta longitud, con un máximo de 280 caracteres (originalmente 140), llamados tuits o tweets, que se muestran en la página principal del usuario. Los usuarios pueden suscribirse a los tweets de otros usuarios, a esto se le llama "seguir" y a los usuarios abonados se les llama seguidores, (followers) y a veces tweeps (Twitter + peeps, seguidores novatos que aún no han hecho muchos tuits). Por defecto, los mensajes son públicos, pudiendo difundirse privadamente mostrándolos únicamente a unos seguidores determinados. Los usuarios pueden twitear desde la web del servicio, con aplicaciones oficiales externas (como para teléfonos inteligentes), o mediante el Servicio de mensajes cortos (SMS) disponible en ciertos países. Si bien el servicio es gratis, acceder a él vía SMS soportar tarifas fijadas por el proveedor de telefonía móvil.

3.0.1. ***API de búsqueda en Twitter.***

Twitter proporciona API de búsqueda de tweets, clasificada en tres métodos:

- ✓ API de búsqueda estándar.

  Devuelve una colección de Tweets relevantes que coinciden con una consulta especificada.

  Tenga en cuenta que el servicio de búsqueda de Twitter y, por extensión, la API de búsqueda no pretende ser una fuente exhaustiva de Tweets. No todos los Tweets se indexarán o estarán disponibles a través de la interfaz de búsqueda. Este servicio es gratuito, sin aprobación previa.
- ✓ API de búsqueda premium.

  Hay dos puntos finales de API de búsqueda premium:
  - Tweets de búsqueda: punto final de 30 días ⟶ proporciona tweets publicados en los últimos 30 días.
  - Tweets de búsqueda: Punto final de archivo completo ⟶ proporciona Tweets desde 2006, comenzando con el primer Tweet publicado en marzo de 2006.

  Estos puntos finales de API de búsqueda comparten un diseño común y la documentación se aplica a ambos. Es gratuita con aprobación previa.

✓ API de búsqueda empresarial.

Hay dos API de búsqueda empresarial:

- API de búsqueda de 30 días: proporciona tweets publicados en los últimos 30 días.
- API de búsqueda de archivo completo: proporciona tweets a partir de 2006, comenzando con el primer tweet publicado en marzo de 2006.

Estas API de búsqueda comparten un diseño común y la documentación se aplica a ambas. Se debe comprar el derecho de uso de esta API, el precio varía según el interés de búsqueda.

3.0.2. ***Regla de búsqueda en Twitter***.

Las reglas se componen de una o más "cláusula" es una palabra clave, una frase exacta o uno de los muchos operadores.

Se puede combinar varias cláusulas con lógica "y" y "o".

- "Y" la lógica se especifica con un espacio entre cláusulas.
- "O" la lógica se especifica con una mayúscula O.

En caso del operador Premium (usado en este proyecto), cada regla puede tener hasta 2048 caracteres de largo sin límites en el número de cláusulas positivas (cosas que desea combinar o filtrar) y cláusulas negativas (cosas que desea excluir y no coincide).

Normas de construcción

Si desea realizar una búsqueda que contenga la palabra Colombianos y Venezolanos sin importar el orden en que aparezcan en el texto. Use el operador lógico "Y"(espacio), así la búsqueda será: Colombianos Venezolanos.

Por otro lado si desea realizar la búsqueda de alguna de las palabras Colombianos o Venezolanos, use el operador "OR" como sigue: Colombianos OR venezolanos.

Ahora si desea excluir palabras puede usar el operador lógico "NO". Por ejemplo buscar la palabra Venezolanos, pero que no contenga colombianos, se escribirá, Venezolanos -Colombianos

Ejemplo de una regla premium

En el ejemplo que sigue a continuación se utiliza los paréntesis para agrupar reglas y lo operadores premium point_radius (latitud, longitud y radio en millas) y el operador premium lang que especificar el lenguaje en español.

Ejemplo .1. (feliz OR fiesta) (vacaciones OR casa OR noche vieja) point_radius: [-105.27346517 40.01924738 10.0mi] lang: es - (cumpleaños OR democrático OR republicano)

Definición de reglas de filtrado en Python.

Explicaremos en cuatros pasos la conexión y la definición de reglas en Python, mediante el uso de la librería **searchtweets.**

- **Paso 1:** Verificación de las credenciales de conexión a la API de Twitter con la función load_credentials().
- **Paso 2:** Definir la regla de filtrado, función gen_rule_payload("regla", results_per_call = 100).
- **Paso 3:** Conexión a la API de Twitter collect_results(rule, max_results = 100, result_stream_args).
- **Paso 4:** Realizamos la consulta a la API, con la función, ResultStream(rule_payload, max_results = 500, max_pages = 1, ∗∗ premium_search_args).

En conclusión, mostramos un resumen de las reglas que hemos usado en este proyecto, para realizar las pruebas que permitieran desarrollar la técnica del método de búsqueda.

✓ Hemos definido regla usando la función gen_rule_payload() de la libreria searchtweets de Python (API Premium).

X No hemos definido reglas en formato JSON de la forma {"valor": "insert_rule_here"}

✓ Se empleó la función filter() de la librería tweepy de Python para filtrado en tiempo real.

✓ Se definió reglas de la forma https://stream.twitter.com/1.1/statuses/filter.json?track=twitter para el filtrado en tiempo real.

Para más información puede visitar la documentación de búsqueda de Twitter [1].

Bibliografía

[1] Documentación de Twitter, https://developer.twitter.com. **Consultada en mayo de 2018.**

[2] mathworks, https://es.mathworks.com/discovery/machine-learning.html. **Consultada en mayo de 2018.**

[3] openmind, https://www.bbvaopenmind.com/que-es-el-aprendijaze-profundo/. **Consultada en mayo de 2018.**

[4] Blei, David M., Andrew Y. Ng, and Michael I. Jordan, Latent Dirichlet Allocation. Journal of Machine Learning Research 3:993-1022, 2003.

[5] B. Pang and L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, Vol 2, No 1-2, pp 1-135, 2008.

[6] C.C. Aggarwal and C.X. Zhai (eds.), Mining Text Data, DOI 10.1007/978-1-4614-3223-4_1, Springer Science+Business Media, LLC 2012.

[7] Mitodru Niyogi and Asim Kumar Pal, Discovering conversational topics and emotions associated with Demonetization tweets in India, arXiv: 1711.04115vl, 2017.

[8] Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press,http://www.deeplearningbook.org, 2016.

[9] Alkire, S. and Santos, M.E., "Acute Multidimensional Poverty: A New Index for Developing Countries,.ºPHI Working Papers 38, University of Oxford. July 2010.

[10] Multidimensional Poverty Peer Network (MPPN), https://www.mppn.org/es/pobreza-multidimensional/. **Consultada en mayo de 2018.**

[11] Rivas A. Sherezade, Minería de texto empleando locación latente de Dirichlet y Aprendizaje Profundo, Universidad Central de Venezuela, Facultad de Ciencias, Postgrado en Modelos Aleatorios, 2016.