UNIVERSIDAD DE MEDELLIN	Reconocimiento de patrones I y II Informe sobre el Análisis de sentimientos	Proyecto de aula. Profesor: Antonio Jesús Tamayo Estudiantes: Luis Eduardo Palacio y Susana Sepúlveda Madrid

En este informe se dará un reporte de dos artículos muy interesantes sobre la minería de textos y análisis de sentimientos, se describirá para cada uno el problema que fue resuelto gracias a sus estudios y como fueron usados los tipos de caracterización para los textos utilizados, también se describirá la metodóloga de validación e implementación y los resultados obtenidos en su clasificación.

Antes de comenzar debemos recordar que en para tomar una decisión, la mayoría de las personas buscan de otras opiniones para dar soporte a la misma y tomar la mejor decisión de acuerdo a lo que se esté consultando, en la actualidad se presentan muchos de estos casos gracias a los comentarios, calificaciones, numero de likes, entre otros, que se encuentran en cualquier página de internet. Por ejemplo, cuando se desea hacer una compra a través de internet, es muy habitual revisar las opiniones asociadas y la calificación al objeto de la compra de cierto proveedor o tienda.

Minería de textos y análisis de sentimientos en sanidadysalud.com

El material de uso para este análisis es tomado del ámbito de consulta de opiniones en el sector sanitario, donde es habitual consultar las opiniones sobre centros o profesionales sanitarios antes de contratar sus servicios. El dominio seleccionado para realizar el reporte, son las opiniones sobre centros de salud, hospitales, farmacias españolas recogidas en el portal web sanidadysalud.com, se espera que con la recolecta de estas diferentes opiniones se puedan identificar los puntos fuertes y los puntos en los que son más susceptibles de mejorar de dichos centros.

“hipótesis de trabajo principal que, mediante técnicas de minería de textos y análisis de sentimientos se puede construir un sistema automático que clasifique correctamente en negativas y no negativas las opiniones escritas en castellano en sanidadysalud.com sobre centros sanitarios españoles”. Sergio Rincón Garcia

- Descripción de la base de datos usada:

Para obtener las opiniones de los usuarios, el portal dispone de cuestionarios de satisfacción o encuestas diferentes sobre diferentes aspectos de los centros de servicios sanitarios, que son complementados libremente. El cuestionario está compuesto por preguntas (socio-demográficas, indicadores de calidad del servicio recibido) y campo de texto para expresar una opinión abierta.
Los datos disponibles abarcan desde el 2010 hasta el 3 de febrero del 2016, por tanto se dividen en base al año el conjunto de datos para entrenar los modelos conseguidos y los textos para evaluar serán desde el año 2015- 2016. Para este sistema se usan las mismas métricas explicadas en el aula de clase que son (Ocurrencia, sensibilidad, especificidad, Medida F de Beta, tasa verdaderos positivos, tasa de falsos positivos, Área balo la curva).

-Técnicas seleccionadas para el modelado:

Como el propósito principal es detectar las opiniones negativas y las no negativas, cualquier opinión que contenga al menos una valoración negativa dentro de una positiva, se considera negativa. Por ello el análisis se realiza a nivel de frase, para que cada opinión sea validada de manera correcta se aplica previamente ciertas reglas, por ejemplo: que este escrito al castellano, que no contengan spam. Luego de ser aplicado y seleccionar las opiniones, teniendo en cuenta también la relevancia, se decide usar el modelado de Naive Bayes y SVM debido a la simplicidad y rapidez de la ejecución, para esto utilizaron

librerías NLTK y Scikit-learn de Python. Los diferentes parámetros de configuración de los algoritmos se decidieron por medio de una validación cruzada de 5 iteraciones y un Grid exhaustivo de búsqueda.

-Implementación:

La extracción de información de los textos para crear el conjunto con los que se alinearan a la técnica de modelado es transformando cada texto en un vector de palabras, luego de tener los datos debidamente organizados donde V es el tamaño del vocabulario para cada W(i) toma el valor de [0,1] esto depende de la aparición de la palabra en el texto t. (el corpus de entrada con tamaño M se transforma en una matriz de dimensión N x M, donde N será el número de características (features) y M el número de observaciones). Para caracterizar los textos, usaron que el vector de palabras tomara valores binarios, en función de si el termino aparece o no en el texto, existen palabras comunes que hacen mucho ruido en la construcción del modelo, por lo tanto, desarrollaron medidas como TF-IDF, que como vimos en el aula de clase (Normaliza ese valor TF con la inversa de la frecuencia de aparición de ese termino en todo el corpus), para 3.359 textos.

-Resultados obtenidos en la clasificación:

Para la cantidad de textos que se tenían se realizaron 4 iteraciones, repetida 20 veces, para evitar en la medida de lo posible el sobreajuste y facilitar la comparación de los modelos obtenidos. Además, para minimizar el problema del gran desequilibrio de las clases objetivo, se utilizó estratificación en el diseño de dicha validación cruzada.

Nota: Para este artículo fueron utilizados gráficos del tipo de diagramas de cajas, su justificación es que para ver el balance sesgo-varianza de los modelos conseguidos es muy ideal. Dado que son bastantes pruebas, en este reporte solo mencionaremos el resultado del artículo de estudio.

"Para este clasificador, los mejores modelos obtenidos se consiguieron eliminando las palabras comunes (SW), reduciendo los términos a su raíz (ST) y seleccionando las mejores características con el test de la χ2 (todo ello con NaiveBayesNLTK)". Sergio Rincón García.

Análisis: Se observa que el objetivo de mercado había sido cumplido ya que el AUC se había marcado como objetivo en 0.9 y los resultados de este análisis fueron de 0.9262, aunque no se utilizó el mejor modelo de validación cruzada, la diferencia o margen de error del problema vino siendo demasiado pequeño, como para generar un ruido.

Tomado de:
https://eprints.ucm.es/39524/1/memoriaTFM_sergio_rincon_garcia.pdf

Análisis de sentimientos en Twitter

El material de uso para este análisis es tomado de la propagación de los comentarios en twitter, específicamente en la cantidad de Twitts, respecto a las elecciones electorales en los EE.UU entre Obama y Romney. El dominio seleccionado para realizar este análisis, son los valores y la visualización que arroja *The New York Times,* Prediciendo los resultados de las elecciones electorales en dicho país en el 2012.

- Descripción de la base de datos usada:

Para este estudio se tomó la red social Twitter, que permite enviar mensajes cortos de 140 caracteres comúnmente llamados twitts. Twitter cuenta con más de 332 millones de usuarios que generan más de 500 millones de tweets al día compartiendo sus puntos de vista y opiniones con sus familiares, conocidos y seguidores. Los usuarios pueden agrupar tweets por tema o por tipo utilizando hashtags - palabras o frases con el prefijo "#". El símbolo "@" acompañado del nombre de un usuario es utilizado para mencionar o responder a dicho usuario. Lo que facilita aún más la búsqueda de un tema en específico.

-Técnicas seleccionadas para el modelado:

El propósito principal es detectar opiniones negativas o positivas respecto a las opiniones de los votantes ciudadanos de los EE.UU, Por ello el análisis se realiza a nivel de frase, para que cada opinión sea validada de manera correcta se aplica previamente ciertas reglas, por ejemplo: Para este análisis debe de ser en ingles los textos, que no información poco relevante. Luego de ser aplicado y seleccionar las opiniones, teniendo en cuenta también la relevancia, se decide usar el modelado de Naive Bayes, para esto utilizaron librerías (Scikit-learn de Python).

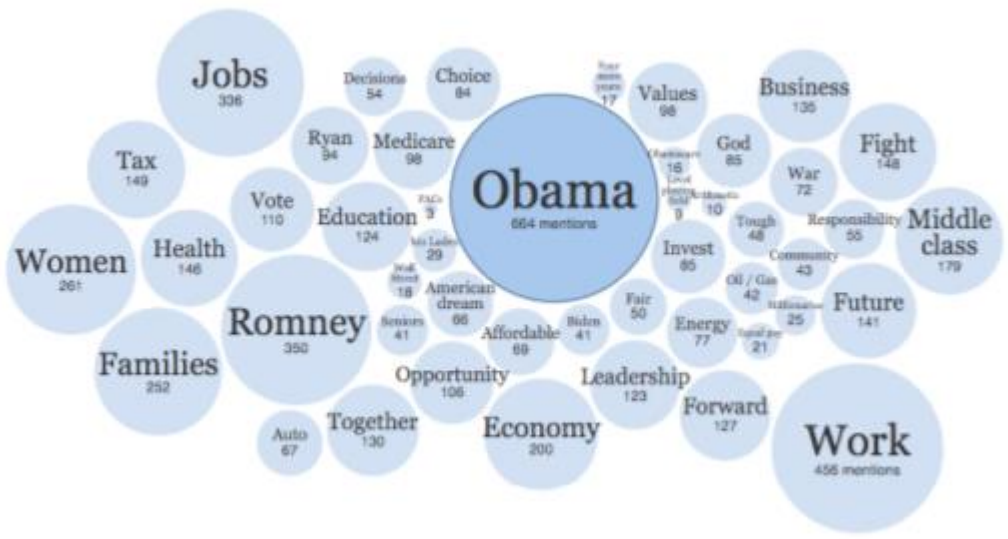
-Implementación:
Se implementa el mismo método que el artículo anterior, el único diferenciador es que el vector de palabras está compuesto por twitts, es decir, un twist hace referencia a un vector de palabras que luego de tener los datos debidamente organizados, para cada W(i) se toma el valor de 0 o 1, esto depende de la aparición de una palabra en el texto, luego de hacer la transformación de la matriz, se toman los valores binarios, también pasa lo mismo que como el ejemplo anterior, podemos tener conectores o palabras de las cuales nos pueden afectar el ruido en la construcción de nuestro modelo, por lo tanto, se pueden llegar a desarrollar medidas como TF-IDF, para este ejemplo en específico no alcanzan a mencionar mucho sobre la finalización de la implementación, ya que no solo se enfocan en twitts.

-Resultados obtenidos en la clasificación:

Para la implementación esta fue la información que arrojo el estudio y fue analizado de la siguiente forma:

Idioma: Inglés, Tweets: 122.443, Tweets únicos: 38.055, Palabras en tweets únicos: 406.311, Palabras únicas: 72.792, Palabras que se repiten en más de 100 tweets: 452, Palabras que se repiten en más de 1000 tweets: 26

Análisis:
En la nube de palabras, cada burbuja contiene un término frecuente y el número de veces que esta se puede encontrar en diferentes discursos, el tamaño de cada burbuja es relativo a la cantidad de menciones del término que representa.



Tomado de:
https://rdu.unc.edu.ar/bitstream/handle/11086/3751/Becerra%202016_analisis-de-sentimiento.pdf?sequence=1