Análisis estadístico de texto

Clasificación de texto con R

Ramirez Hurtado Tito Jaime
202013426@urp.edu.pe

Caso práctico

El conjunto de datos consta de 1349 tuits acompañados de su nombre de usuario e identificador de tuit. Estos tuits fueron obtenidos el 9 de Abril del 2018.

El conjunto de datos contienen tuits que pertenecen a cuatro cuentas, mezclados con tuits de multiples usuarios.

@lopezobrador - Andrés Manuel Lopez Obrador, candidato a la presidencia de México.

@UNAM MX - Universidad Nacional Autónoma de México, cuenta institucional.

@CMLL OFICIAL - Consejo Mundial de Lucha Libre, promoción de lucha libre de México.

@MSFTMexico - Microsoft México, cuenta corporativa.

Variables del conjunto de datos:

Variables	**Descripcion**
status_id	Código del registro
screen_name	Usuario quien realizó el tuit
text	Comentario/tuit que realiza el usuario.

1. Preprocesamiento de Texto

Antes de empezar a cargar los datos, es necesario contar con las siguiente librerías **tidyvers,tidytext,naivebayes,tm,caret dplyr, Rcpp** .

Se realiza la carga de los datos en memoria y se visualiza una muestra de las variables del conjunto de datos tuits.csv .

```
# A tibble: 6 x 4
  status_id screen_name   text                                                    ID
      <dbl> <chr>         <chr>                                                <int>
1   8.32e17 lopezobrador_ Josefina Vázquez Mota debe informar qué hizo con los mil millones de pes~     1
2   8.33e17 lopezobrador_ Ya ni la burla perdonan: bajaron 2 centavos la gasolina. En Guatemala no~     2
3   8.34e17 lopezobrador_ Martín Moreno no votará por mí. Comprendo. Es un mal escritor dedicado a~     3
4   8.34e17 lopezobrador_ En Chicago dije que iremos a Nueva York (ONU) y a Washington (CIDH) el 1~     4
5   8.34e17 lopezobrador_ Entérate y apoya con tu firma la denuncia contra las órdenes de Donald T~     5
6   8.35e17 lopezobrador_ Qué república ni qué ocho cuartos, es la monarquía de la moronga azul. C~     6
```

Es importante conocer la cantidad de tuits por Autor, la siguiente figura 1 muestra la cantidad de tuits por **Autor** en todo el conjunto de datos; hasta este momento no realizó ningun tratatiento a los datos. Tan sólo se realiza un conteo a nivel de Autor.

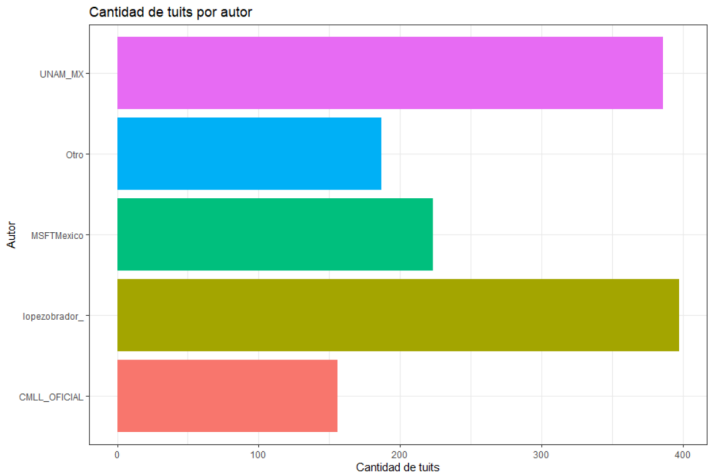

Figura 1: Cantidad de Tuits por Autor

1.1. Limpieza y Tokenización

Se realiza un proceso de limpieza de los stopwords y realiza la tokenización para analizar las palabras a nivel de Autor.

La siguiente figura 2 muestra una comparativa entre la media, promedio de la cantidad de tuits por **Autor**, esto considerando ya haber quitado los stopwords.

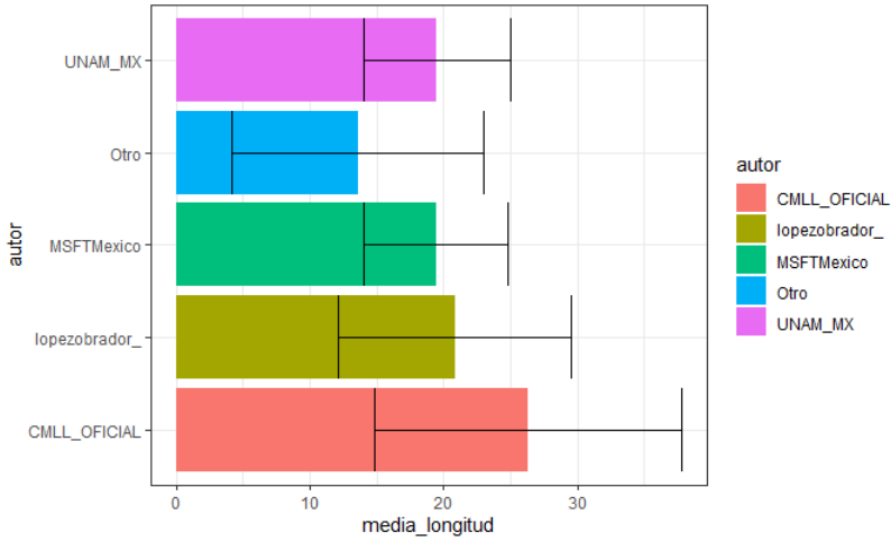

Figura 2: Digramas de las media longitud

También es importante conocer el top n de palabras más utilizadas en los tuits por Autor.

La siguiente figura 3 muestra la información de las palabras más utilizads por los autores de donde se puede ver que lo mas importante son **arena,mexico,microsoft,video,gf**.

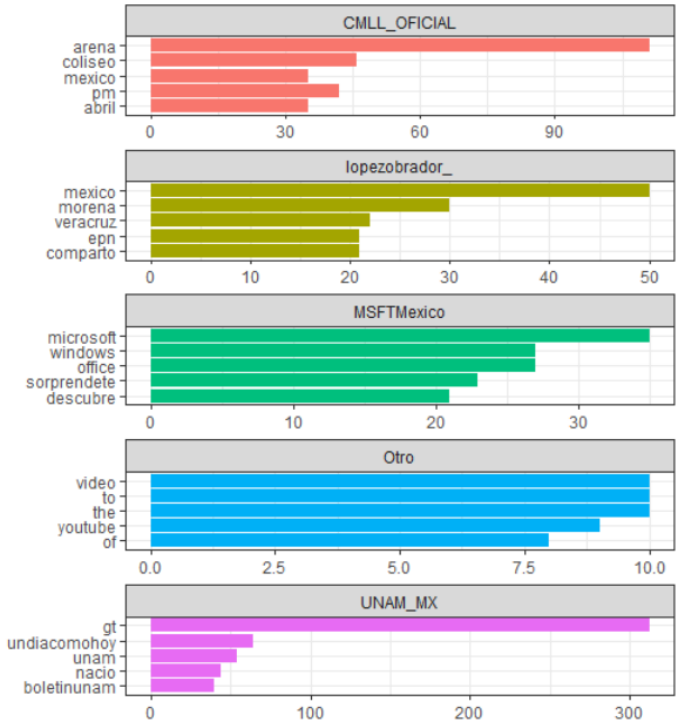

Figura 3: Top 5 de palabras mas utilizadas por el Autor

También es importante conocer la frecuencia de las palabras más utilizadas por un autor; considerando el autor **MSFTMexico** se realiza un análisis de las palabras más utilizas y en la figura 4 se puede observar que las palabras como **microsoft,windows,soprendete, descubre, office, microsoftsessions** son algunas de las que más utiliza.

Figura 4: Palabras más utilizadas por e autor MSFTMexico

Otro punto de análisis importante es comparar entre autores para ver que tanto coinciden en el uso de las palabras. En esta comparación de la figura 5 se puede ver que los autores tanto MSFTMexico diferen porque son autores que tiene enfoques muy distintos en sus contenidos de los tuits.

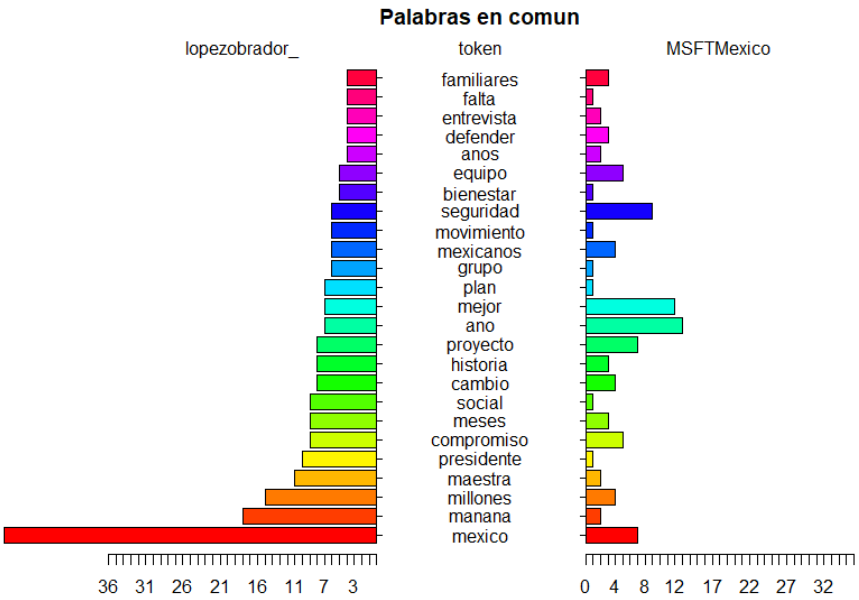

Figura 5: Palabras comunes entre los autores de lopezobrador y MSFTMexico

1.2. Análisis de bigramas

Un bigrama o digrama es un grupos de dos letras, dos silabas, o dos palabras. Los bigramas son utilizados comunmente como base para el simple analisis estadistico de texto. Se utilizan en uno de los más exitosos modelos de lenguaje para el reconocimiento de voz. Se trata de un caso especial del N-grama.

Gracias a los bigramas se puede entender algunas relaciones de las palabras utilizadas en el contexto de los tuits por los autores. Uno consideración a esto es retirar los bigramas que tiene o son generados de una combinación con los stopwords.

La siguiente Figura 6 muestra la relación digramas que mas se utilizan en los tuits.

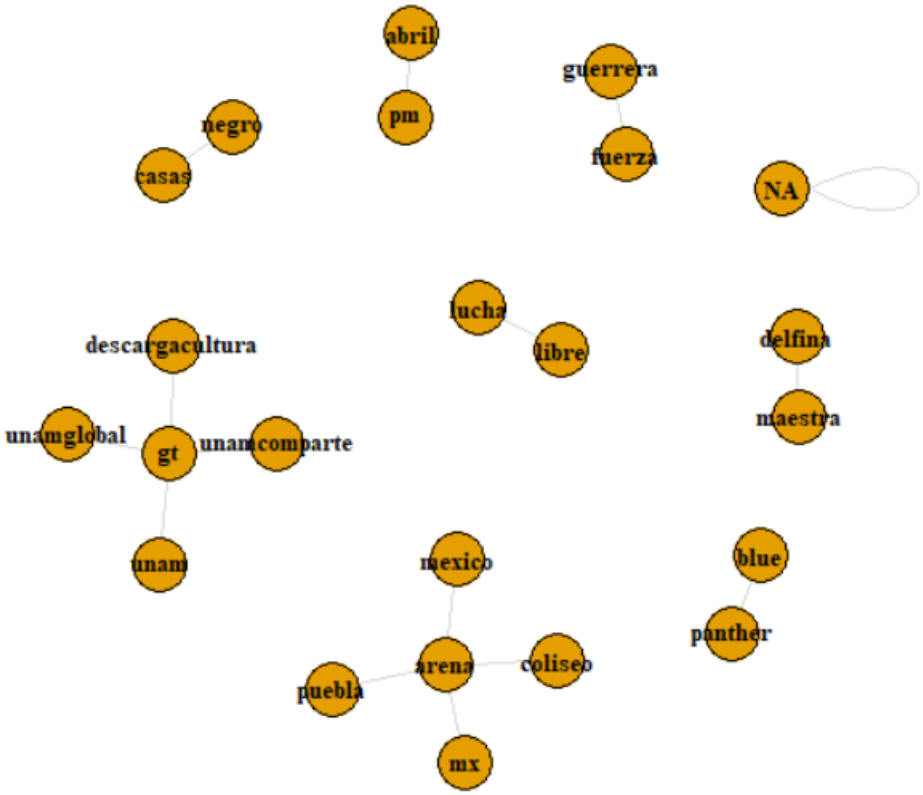

Figura 6: Digramas de las palabras de los tuits

2. Modelos - Clasificación

2.1. Naive bayes

En teoría de la probabilidad y minería de datos, un clasificador Naive Bayes es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de naive, es decir, ingenuo.

Formula:

$$p\left(y|x\right) = \frac{p\left(x|y\right)p\left(y\right)}{p\left(x\right)}$$

Se realiza una partición de 70 % - 30 % y se aplica el algoritmo naive bayes considerando la relación del autor lopezobrador_.

La siguiente tabla 1 muestra la validación cruzada entre el autor lopezobrador_ vs otros.

	lopezobrador_	Otro
lopezobrador_	105	20
Otro	16	258

Cuadro 1: Validación cruzada

De los resultados se obtiene que el Accuracy es de 0,9098 el cual se va comparar mas adelante con los resultados del SVM.

```
              Accuracy : 0.9098
                95% CI : (0.8773, 0.936)
    No Information Rate : 0.6967
    P-Value [Acc > NIR] : <2e-16

                 Kappa : 0.7885

 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 0.8678
            Specificity : 0.9281
         Pos Pred Value : 0.8400
         Neg Pred Value : 0.9416
             Prevalence : 0.3033
         Detection Rate : 0.2632
   Detection Prevalence : 0.3133
      Balanced Accuracy : 0.8979

       'Positive' Class : lopezobrador_
```

2.2. SVM

Este algoritmo tiene muy buenos resultados para diversas tareas de procesamiento de lenguaje natural (NLP) como por ejemplo clasificación de textos.

El algoritmo SVM representa el documento de texto como un vector donde la dimensión es el número de palabras distintas.

Si el tamaño del documento es grande, entonces las dimensiones son enormes, si esto ocurre en la clasificación de texto se tendrá un alto coste computacional. Para lograr un mejor desempeño (un incremento entre el 1 % a 5 %), se deben evaluar en diferentes niveles los parámetros (Aliwy Ameer, 2017).

	lopezobrador_	Otro
lopezobrador_	100	15
Otro	13	277

Cuadro 2: Validación cruzada

```
              Accuracy : 0.9309
                95% CI : (0.9016, 0.9536)
    No Information Rate : 0.721
    P-Value [Acc > NIR] : <0.0000000000000002

                 Kappa : 0.8291

 Mcnemar's Test P-Value : 0.8501

            Sensitivity : 0.8850
            Specificity : 0.9486
```

```
        Pos Pred Value : 0.8696
        Neg Pred Value : 0.9552
            Prevalence : 0.2790
        Detection Rate : 0.2469
  Detection Prevalence : 0.2840
     Balanced Accuracy : 0.9168

      'Positive' Class : lopezobrador_
```

3. Conclusiones

En resumen se puede observar que comparando el Accuracy entre Naive bayes 0,9098 and SVM 0,9309 , el mejor modelo que se ajusta a la clasificación es el de Naive bayes.