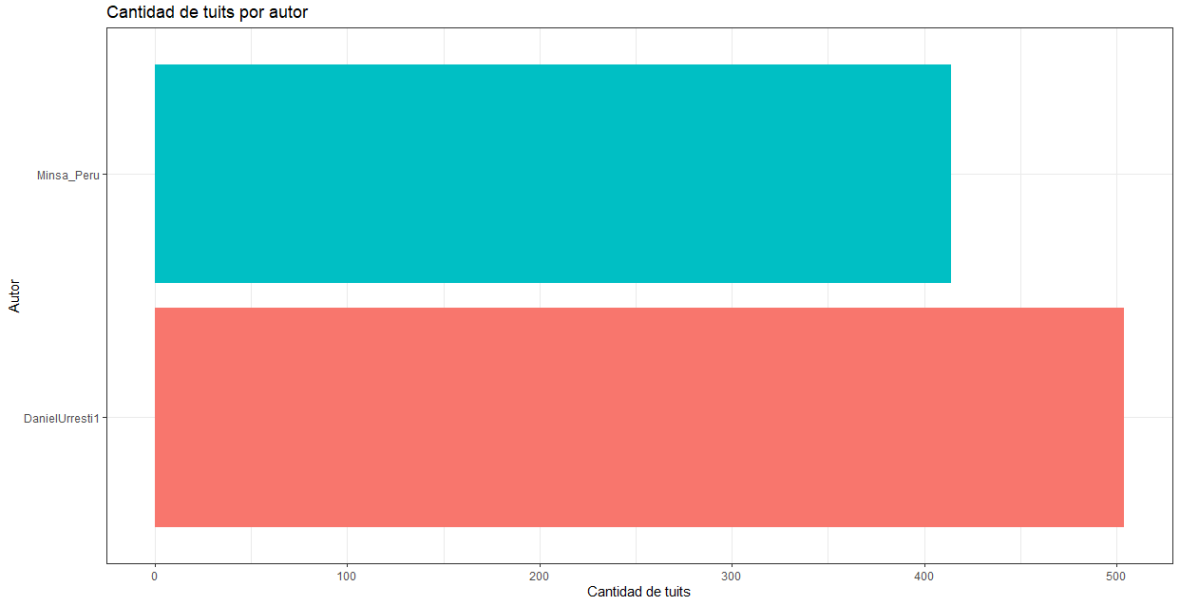
Maestría de Ciencia de Datos **Curso: Análisis de Estadístico de Texto**

Alumno: Víctor Melchor Espinoza

Código: 202013441

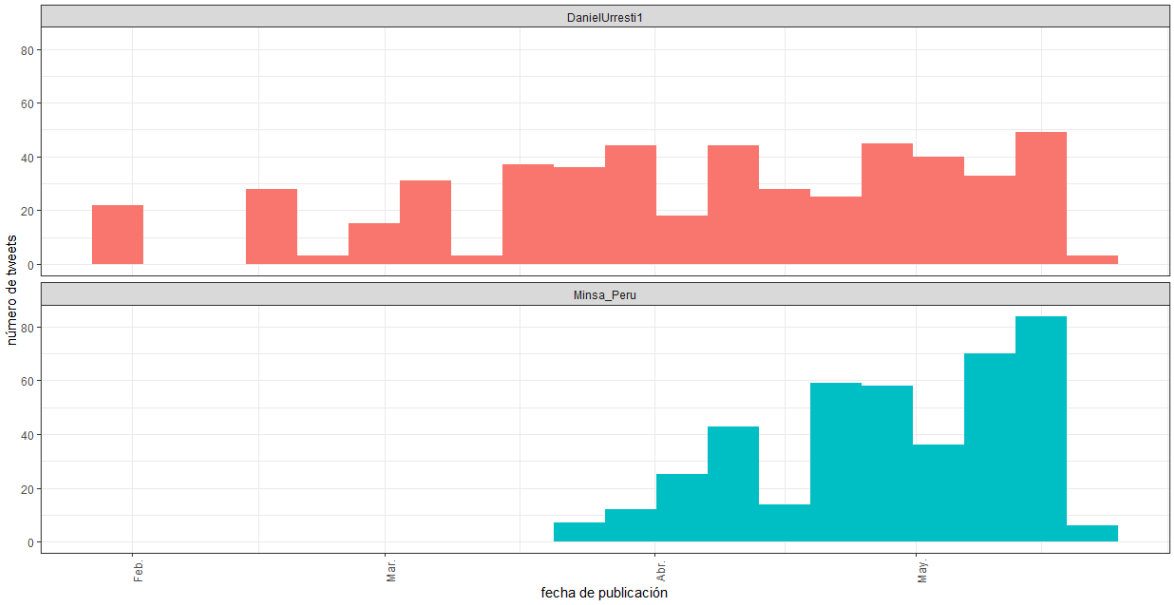
I. Análisis Exploratorio

Cantidad de Tweets por autor



AUTOR	numero_label
<chr>	*<int>*
1 DanielUrresti1	504
2 Minsa_Peru	414

Daniel Urresti se ha mantenido más activo por Twitter desde Enero mientras que Minsa recién a fines de Marzo.

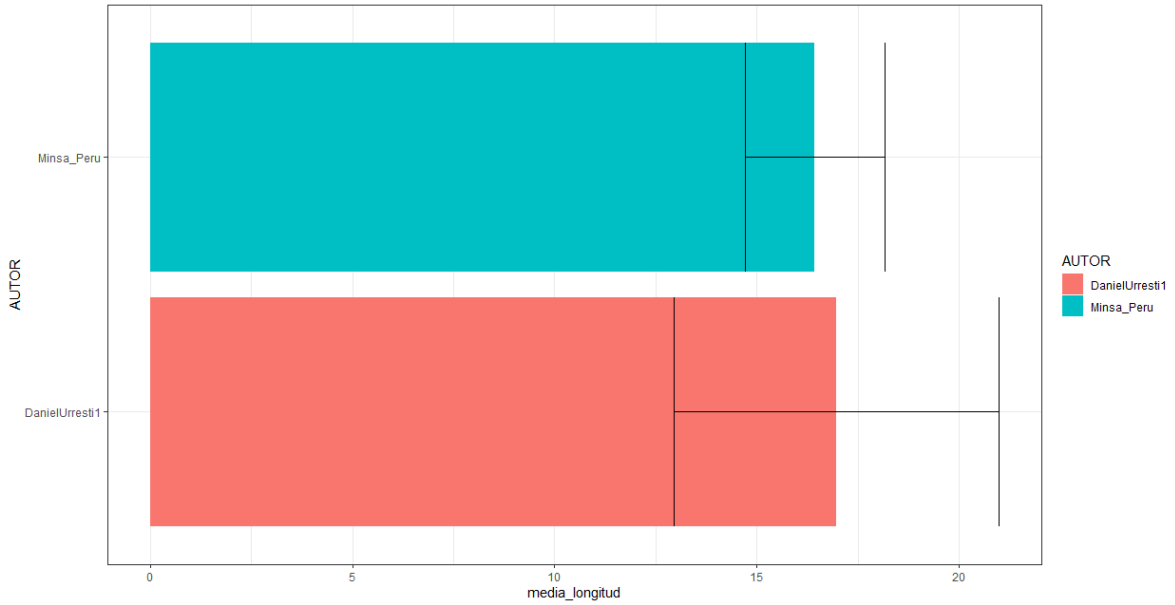


Aquí vemos el total de palabras por autor

```
  AUTOR              n
1 DanielUrresti1  8532
2 Minsa_Peru      6804
```

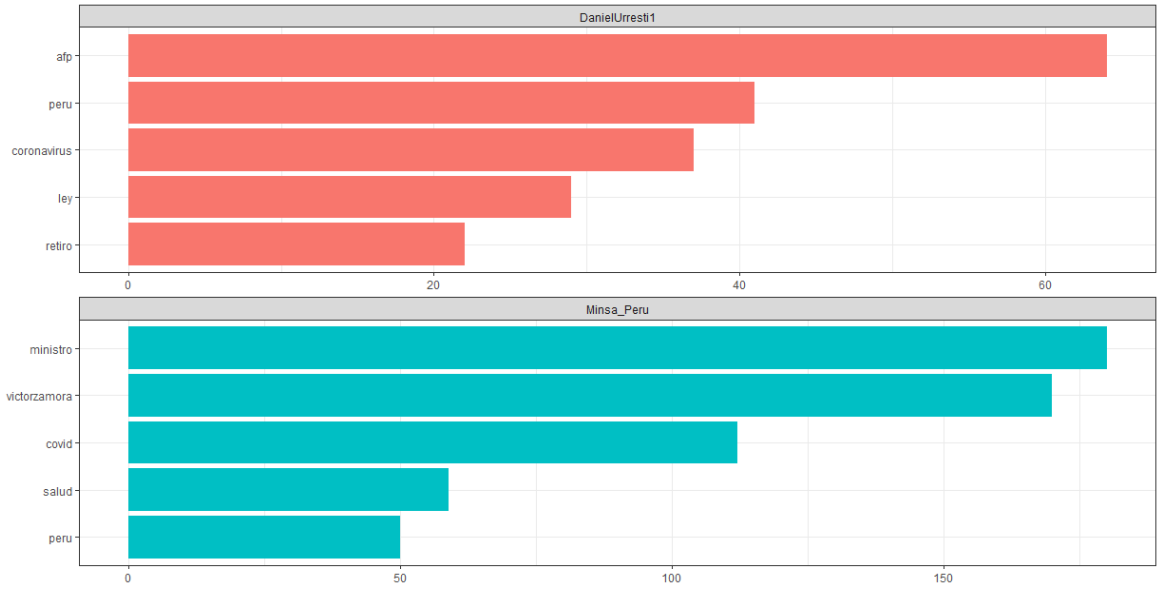
Al analizar la longitud media , vemos que su media es muy cercana en ambos, sin embargo Urresti tiene mayor varianza

```
  AUTOR          media_longitud sd_longitud
  <chr>                   <dbl>       <dbl>
1 DanielUrresti1           17.0        4.02
2 Minsa_Peru               16.4        1.73
```



II. Palabras más usadas por Autor

Urresti usa más afp, Perú, mientras que MINSA, ministro VictorZamora, covid



III. Nube de Palabras

Daniel Urresti

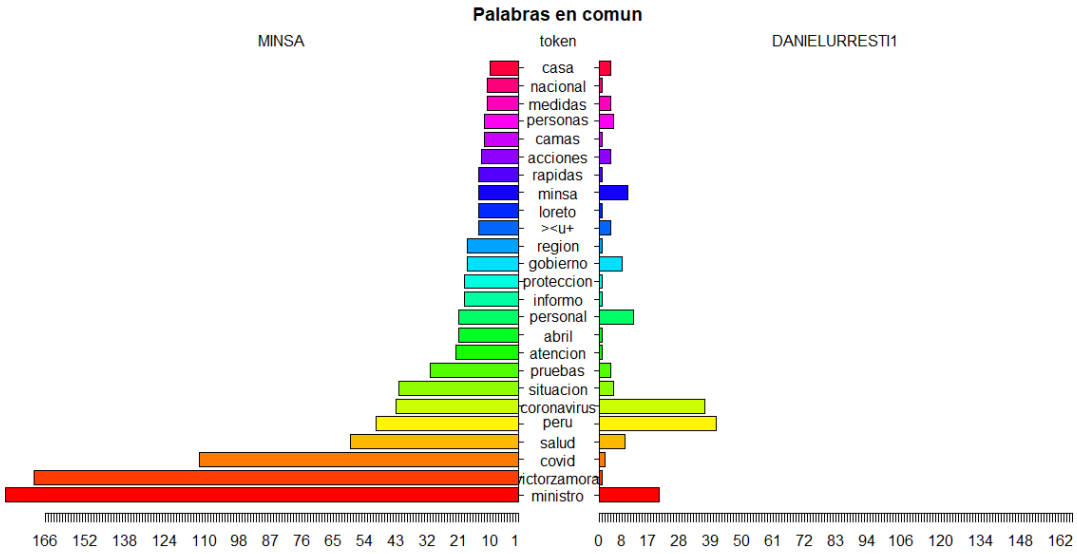
En este caso el discurso de Urresti era el proponer un proyecto de ley acerca del retiro de dinero de AFP por motivo del coronavirus en tiempo de cuarentena.

MINSA

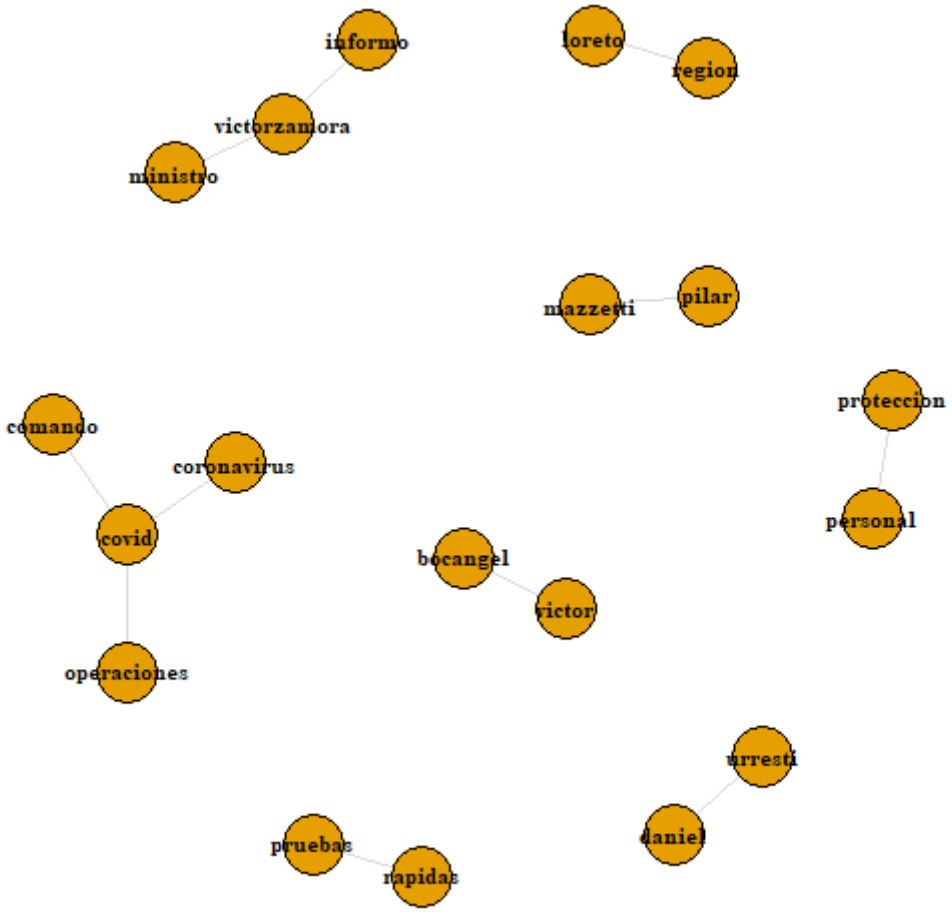
Por otro lado en el MINSA el ministro daba las noticias diarias acerca de la situación de salud que afectaba a los peruanos por el COVID.

IV. Palabras en Común

En el gráfico de Pirámide se observa que el MINSA usa más las palabras Ministro, VictorZamora y covid. Mientras que Urresti usa Coronavirus, Perú y Ministro.



En el análisis de bigramas se observa el diagrama de red en la que se puede notar la relación entre palabras como Victor Zamora,ministro e informa, Pilar Mazzetti, Daniel urresti, y covid con comando, coronavirus, operaciones



V. Modelado

```
Confusion Matrix and Statistics

                Reference
Prediction       DanielUrresti1 Minsa_Peru
  DanielUrresti1            102         16
  Minsa_Peru                 48        110
```

Para el modelado se usó Naive Bayes sin normalización, SVM con corpus limpio y TDIDF y Naive Bayes normalizado obteniéndose los siguientes resultados:

a) Naive bayes sin normalizacion

Daniel Urresti1

```
> precisiona
[1] 0.8644068
> recalla
[1] 0.68
> F1a
[1] 0.761194
```

b) SVM con corpus limpio y tdidf

Proporción de los grupos train y test

```
DanielUrresti1     Minsa_Peru
     0.5327103      0.4672897
DanielUrresti1     Minsa_Peru
     0.5869565      0.4130435
```

```
                observado
predicho         DanielUrresti1 Minsa_Peru
  DanielUrresti1            152          8
  Minsa_Peru                 10        106
```

```
> precisionb
[1] 0.95
> recallb
[1] 0.9382716
> F1b
[1] 0.9440994
```

c) Naive bayes normalizado

Tabla de Proporciones

```
DanielUrresti1     Minsa_Peru
     0.5327103      0.4672897

DanielUrresti1     Minsa_Peru
     0.5869565      0.4130435
```

```
> precisionc
[1] 0.9741935
> recallc
[1] 0.9320988
> F1c
[1] 0.9526814
```

Finalmente comparamos los indicadores en la siguiente tabla

```
  precisiona precisionb precisionc
1  0.8644068       0.95  0.9741935
> cbind.data.frame(recalla,recallb,recallc)
  recalla   recallb   recallc
1    0.68 0.9382716 0.9320988
> cbind.data.frame(F1a,F1b,F1c)
       F1a       F1b       F1c
1 0.761194 0.9440994 0.9526814
```

Al comparar los modelos se constata que Naive Bayes tiene mejor precisión(97%), evaluando el **recall,** es muy parecido en ambos casos y en **F1 Score,** Naive Bayes otra vez es superior con un 95%,

Por lo que podemos afirmar que Naive Bayes tiene un mejor comportamiento que SVM.