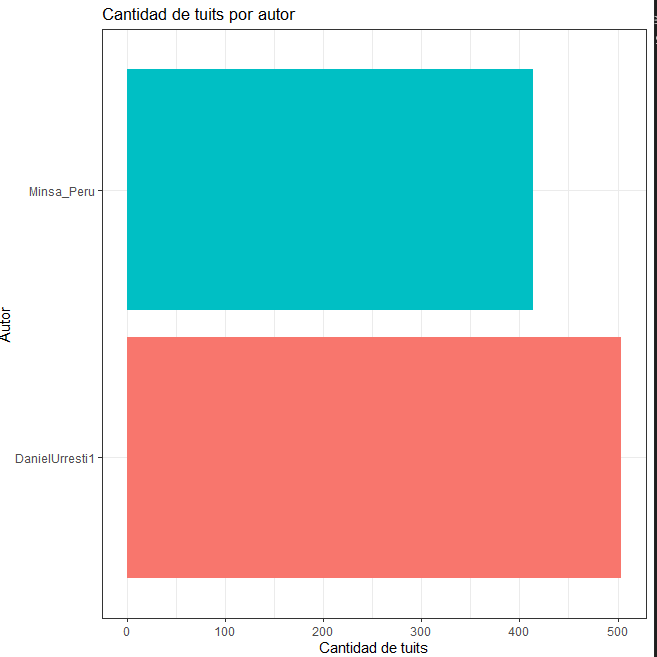
Tarea Sesión 04

Alumno: Miguel Angel Espinoza Carranza

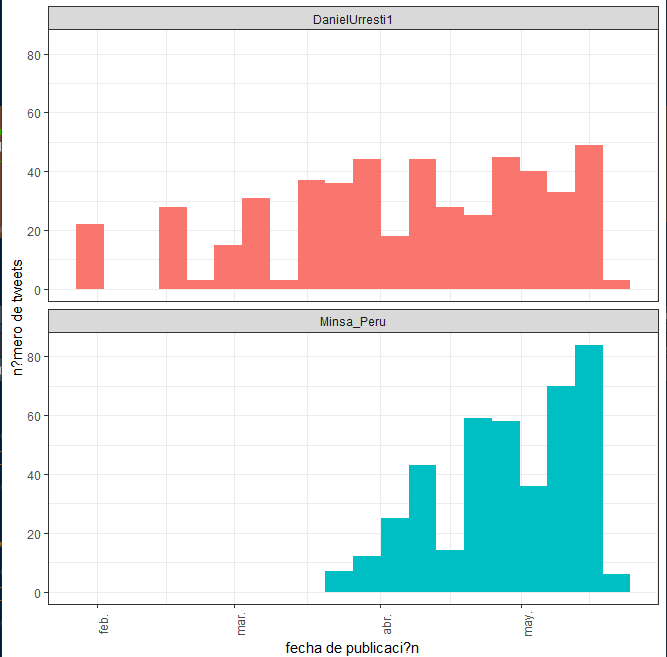
El siguiente trabajo revisará los tweets realizados por las cuentas de Daniel Urresti y la del Ministerio de Salud (Minsa), entonces como primer paso haremos una revisión rápida de estos registros:

**Análisis Exploratorio**

## Cantidad de tweets por autor. ##

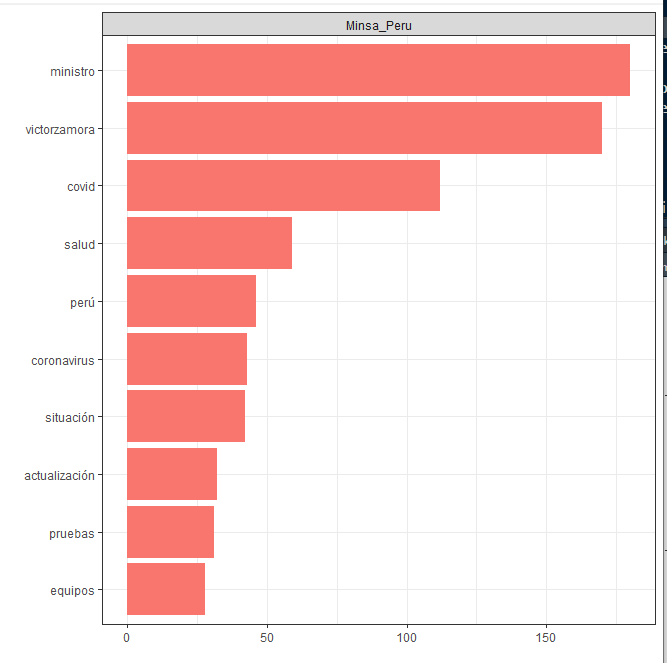
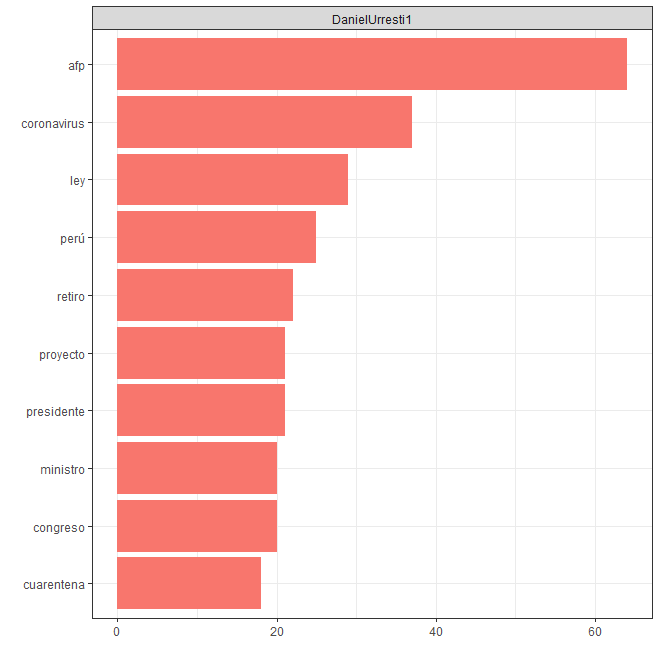


## Cantidad de tweets por autor por fecha de publicación. ##



Como podemos observar ambos tienen una cantidad de tweets de entre 400 y 500, aunque podemos ver también que los tweets del Minsa están centrados entre los meses de abril y junio; mientras que los de Daniel Urresti aparecen también en meses previos a abril.

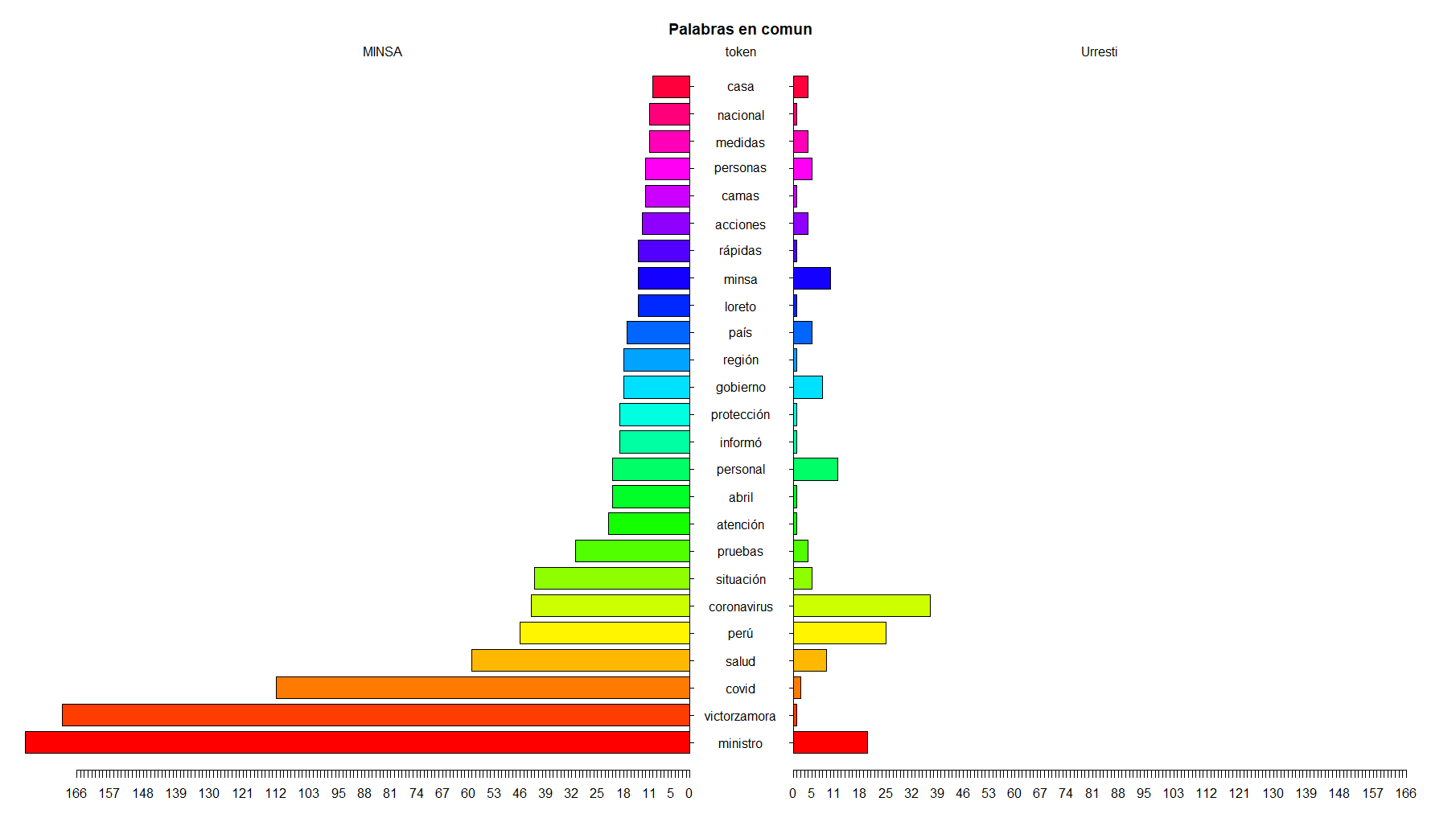
Por otro lado, también se hizo la revisión de las palabras más usadas por cada autor, para el caso del Minsa se tienen palabras relacionadas a la coyuntura del covid, se observa también la palabra victorzamora quien era ministro de salud en esas fechas, mientras que Daniel Urresti tiene entre sus palabras más usadas, temas relacionados a la coyuntura del covid y además del tema relacionado al proyecto de ley de retiro de afp.

Adicional a ello también se presenta la nube de palabras de cada uno de los autores, los cuales permiten ver de una forma más dinámica las palabras más usadas, del lado izquierdo tenemos la información del Minsa, y del lado derecho la de Daniel Urresti:

Como se puede ver, la nube de palabras del Minsa está más relacionado a temas de covid, y aunque Daniel Urresti también toca este tema, se observa que también toca otros temas como el proyecto de retiro de la afp, también se puede observar mención a ministros y al presidente.

Otra comparativa importante a realizar es el análisis de términos comunes usados por ambos autores:



Como era de esperarse los términos en común de estos dos autores está relacionado al manejo de los casos por covid, si vemos más a detalle vemos términos como ministros que es usado en ambos casos; sin embargo, en el caso del Minsa hace referencia a los comunicados realizados por el ministro, mientras que probablemente Daniel Urresti haya usado ese término para realizar alguna crítica a su gestión. También se observa que dentro de las palabras en común tenemos a “Perú”, ya que obviamente estos autores son de ese país, se debería evaluar si retirarlo o no, ya que no aporta mayor contexto.

**Modelos de Clasificación**

También se realizó el modelado usando tres modelos de clasificación: Naive Bayes sin normalización, SVM con corpus limpio y IT-IDF , y Naive Bayes normalizado, obteniéndose los siguientes indicadores:

|  | **Naive Bayes**  **Sin normalización** | **SVM con Corpus Limpio** | **Naive Bayes**  **Normalizado** |
| --- | --- | --- | --- |
| Precisión | 86.4% | 93.1% | 95.5% |
| Recall | 68.0% | 92.0% | 90.7% |
| F1 | 76.1% | 92.5% | 93.0% |

En base a tabla podemos ver que el modelo de Naive Bayes sin normalización tiene los indicadores más bajos, ya aplicando normalización observamos que los indicadores mejoran considerablemente. Si comparamos los modelos de SVM Y NV normalizado podemos ver que tienen indicadores muy parecidos, si nuestro objetivo es el de minimizar los falsos negativos , es decir, guiarnos por el Recall, el modelo escogido sería SVM