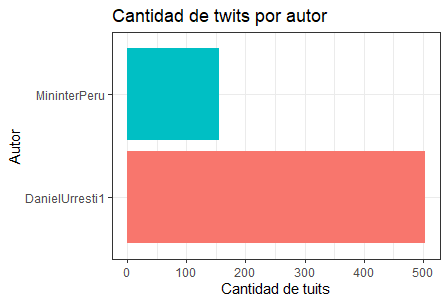
**TAREA Nº 3 - Análisis de Estadístico de Texto**

**ALUMNO:** Danny Canales Quiñonez

**CUENTAS DE USUARIO: @MinterPeru y @DanielUrresti**

**ETAPA 1: Análisis Exploratorio**

Cantidad de twits por cuenta de usuario.



Del gráfico se observa que la mayor cantidad de twits es de la cuenta @DanielUrresti con un total de 504 twits y de la cuenta del @Mininterperu es de 155 twits.

A continuación, se muestra la nube de palabras:

|  |  |
| --- | --- |
| **@DanielUrresti** | **@MininterPeru** |
|  |  |

De los gráficos para cada cuenta se observa que para la cuenta de @DanielUrresti las palabras afp, peru y coronavirus son las que se usaron con mayor frecuencia; mientras que para la cuenta @@ MininterPeru es ministro y gastonrodriguez.

Analizamos las palabras comunes entre las dos cunetas:

Gráfico

Descripción generada automáticamente

Del gráfico se observa que no existe palabras comunes entre ambas cuentas, debido a que son cuentas de distinta jerarquía la primera es de una persona y la segunda es de una institución.

**ETAPA 2: Modelamiento**

A continuación, se utiliza dos modelos para la clasificación de textos: Maquina de Soporte Vectorial (SVM) y de Naive Bayes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algoritmo** | **SVM (BOW)** | **SVM (TF\_IDF)** | **Naive Bayes (BOW)** | **Naive Bayes (TF\_IDF)** |
| **Precisión** | 0.9756 | 0.7742 | 0.9048 | 0.8519 |
| **Recall** | 0.8163 | 0.9796 | 0.7308 | 0.9388 |
| **F score** | 0.8889 | 0.8649 | 0.8085 | 0.8932 |

De la tabla se puede observar que el mejor modelo de clasificación es Naive bayes con la matriz de TF\_IDF puesto que tiene las mejores métricas de clasificación.