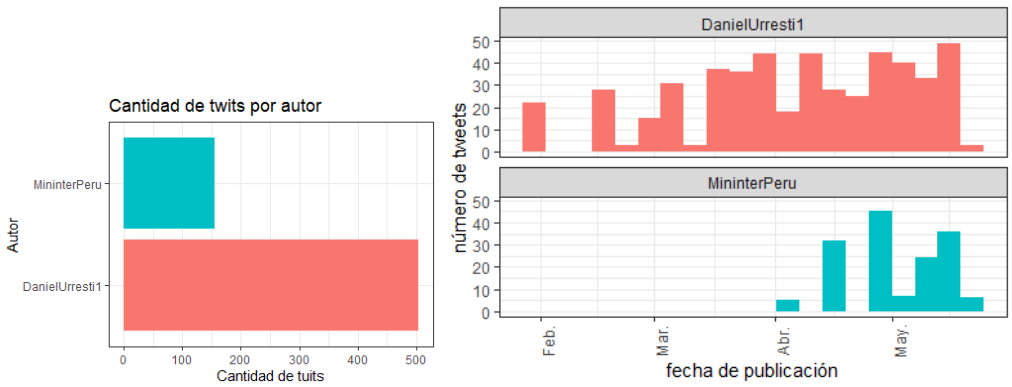
Práctica Nro. 3

Alumno: Esteban Cabala, Jesús Guillermo

Clasificación twits: MinterPeru y DanielUrresti

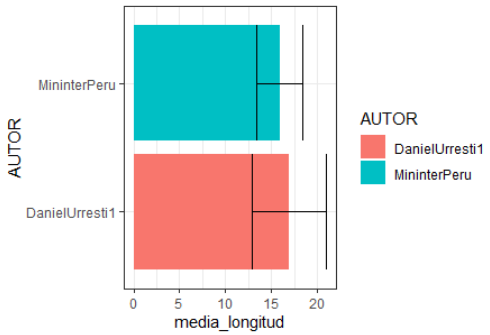
Análisis Exploratorio

1.- Cantidad de Twits



Hay mayor cantidad de twits de DanielUrresti (504) que del ministerio del Perú (155), los twits del miniterPeru. Los del ministerio del interior empezaron a salir desde abril, mientras que los de Urresti se dieron de forma más prolija en el tiempo

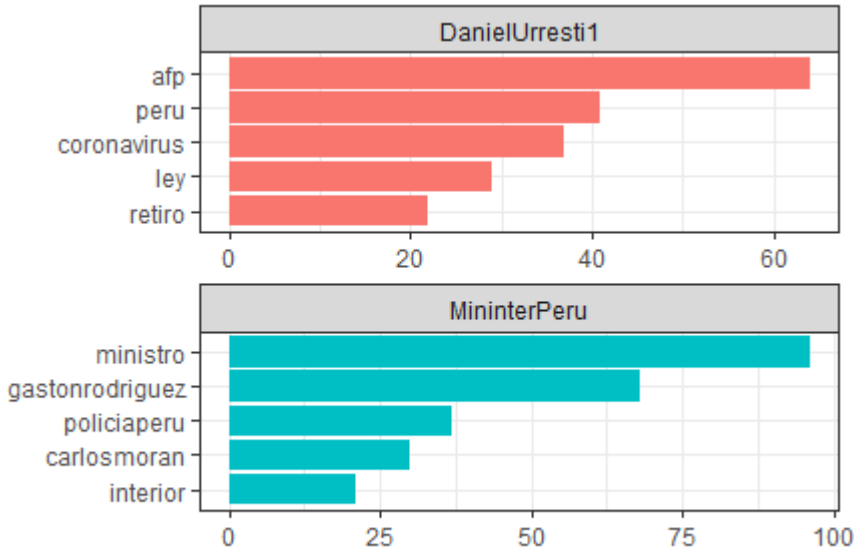
2.- Media de palabras por autor



AUTOR	media_ longitud	sd_ longitud
DanielUrresti1	17	4.02
MininterPeru	15.9	2.55

Al tener más twits, Urresti también tiene más palabras utilizadas, así mismo tiene una mayor longitud de palabras utilizadas en cada twit y una mayor variedad en la longitud de cada twits

3.- Palabras más utilizadas por autor

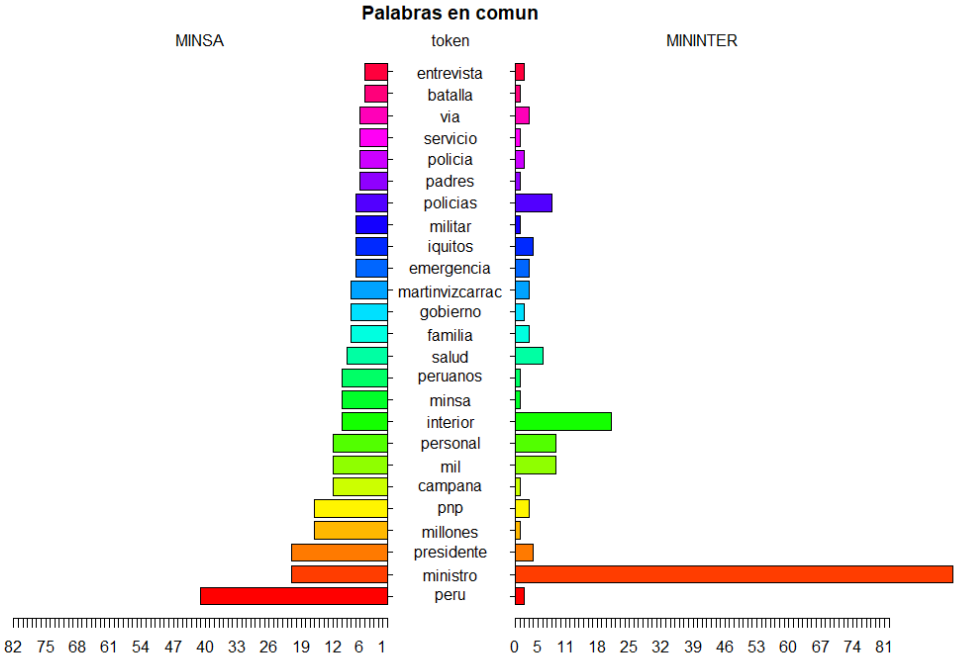


Daniel Urresti | Ministerio del Interior

Las palabras más usadas por Urresti son afp, peru, coronavirus, ley, retiro. Refiriéndose seguro a los temas del retiro de la afp. Mientras que el ministerio del interior son ministro, gatonrodriguez, policiaperu, carlormoran, interior. Ya que se refiere más a citas de comunicaciones oficiales de estos personajes y/o instituciones.

En los wordclouds también evidenciamos estas palabras, y a su vez que Urresti ha usado una mayor variedad léxica que el ministerio del interior

4.- Palabras en común



En las pas 25 palabras en común mas utilizadas podemos ver algunas diferencias, ministro lo usa mucho más el ministerio que Urresti, mientras que Perú lo hace urresti. Hay palabras que las usan de manera similar como son policías, personal, mil.

Modelamiento

Se procedió a realizar el modelamiento, pero son con palabras limpias (normalizadas) es decir con limpieza y sin stopwords

Se hizo 4 modelos en total, 2 con Máquina de Soporte Vectorial (SVM), uno con una Bolsa de Palabras de frecuencias y el otro con el indicador de TF_ITF, así mismo se realizó otros dos modelos semejantes, pero con el algoritmo de Naive Bayes.

Obteniéndose los siguientes resultados (las métricas se basan en la posibilidad de clasificar correctamente el twit para el ministerio del interior):

Matriz de Confusión de los modelos

SVM con BOW (frecuencias)

		Reference	
		DanielUrresti1	Mininter Peru
Prediction	DanielUrresti1	148	9
	Mininter Peru	1	40

SVM con TF_IDF

		Reference	
		DanielUrresti1	Mininter Peru
Prediction	DanielUrresti1	135	1
	Mininter Peru	14	48

NaiveBayes con BOW (frecuencias)

		Reference	
		DanielUrresti1	Mininter Peru
Prediction	DanielUrresti1	142	14
	Mininter Peru	4	38

NaiveBayes con BOW (frecuencias)

		Reference	
		DanielUrresti1	Mininter Peru
Prediction	DanielUrresti1	141	3
	Mininter Peru	8	46

Indicadores de Clasificación de los Modelos (MininterPeru)

	Precision	Recall	F1
SVM con BOW (frecuencias)	0.9756	0.8163	0.8889
SVM con TF_IDF	0.7742	0.9796	0.8649
NaiveBayes con BOW (frecuencias)	0.9048	0.7308	0.8085
NaiveBayes con TF_IDF	0.8519	0.9388	0.8932

El modelo que mejores resultados nos brindó (basándonos en el indicador de F1 score) fue el NaiveBayes con la matriz de TF_IDF, con un F1 score de 0.8932, el modelo de SVM con una bolsa de palabras fue el segundo algoritmo que mejores resultados nos dio.

Conclusiones

Los twits entre el ministerio del Interior y de Daniel Urresti, se diferencian claramente por las palabras utilizadas, Urresti en general comenta más temas, mientras que el ministerio del interior responde a comentarios que puedan hacer los ministros de la cartera.
Es así que se puede hacer una buena clasificación por medio de algoritmos como Naive Bayes y Máquina de Soporte Vectorial, para la clasificación de twits correspondientes al ministerio del interior, el algoritmo de Naive Bayes con una Bolsa de palabras con el indicador de TF_IDF fue el que mejores resultados otorgo (basándonos en la metrica de F1-Score)