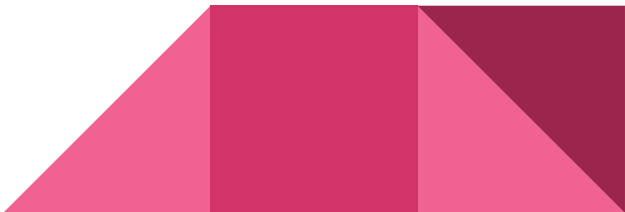


Text mining en social media

Detección automática de género y variedad

¿Qué hemos probado?

- Stop Words adicionales
 - Diferentes rangos de términos frecuentes
 - Vocabulario de bigramas
 - Diferentes tamaños de vocabulario
 - TF_IDF
 - SVM
 - Random Forest
- 

Stop Words

No mejora el resultado



Vocabulario de bigramas

No hemos mejorado el base line

Mejor ajuste género: 0.6



TF_IDF

1. **N** palabras más frecuentes por categoría.
2. Cálculo TF_IDF de la librería TidyText.
3. Vocabulario con las **M** palabras más discriminantes.



TF_IDF por género

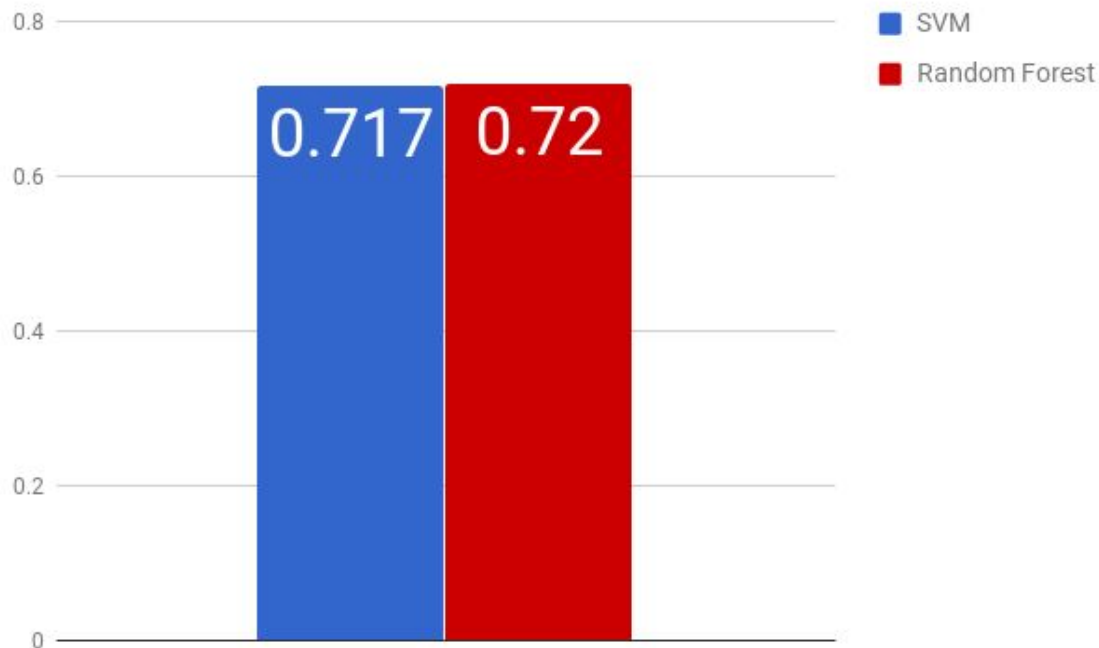
	WORD	FREQ	gender	tf	idf	tf_idf
1	gol	474	male	0.0011256795	0.6931472	0.0007802616
2	vs	430	male	0.0010211860	0.6931472	0.0007078322
3	retweeted	418	male	0.0009926878	0.6931472	0.0006880788
4	whatsapp	397	female	0.0009578359	0.6931472	0.0006639213
5	sola	375	female	0.0009047568	0.6931472	0.0006271297
6	llorar	357	female	0.0008613285	0.6931472	0.0005970274

TF_IDF por variedad

	WORD	FREQ	variety	tf	idf	tf_idf
1	peru	1052	peru	0.008241543	1.94591	0.016037302
2	vos	949	argentina	0.007451261	1.94591	0.014499484
3	bogota	854	colombia	0.006586355	1.94591	0.012816456
4	elcooperante	869	venezuela	0.006378356	1.94591	0.012411708
5	nicolasmaduro	723	venezuela	0.005306734	1.94591	0.010326427
6	epn	584	mexico	0.004691554	1.94591	0.009129343

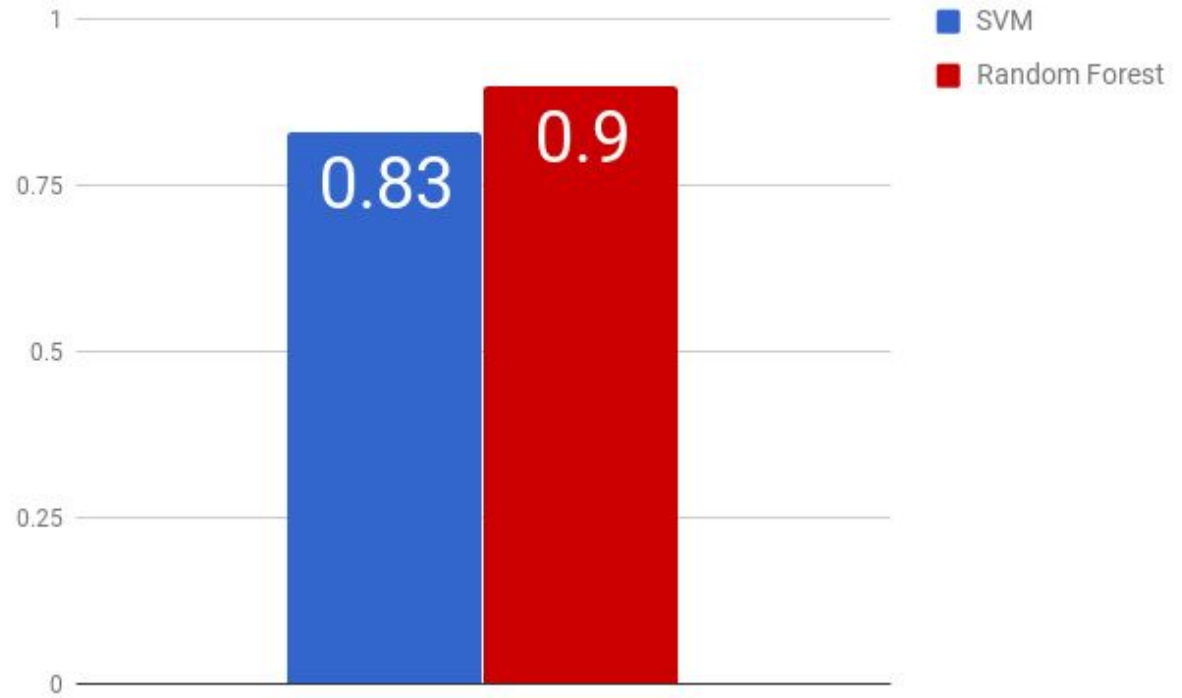
SVM vs Random Forest

- Gender
- TF_IDF
- 300 palabras



SVM vs Random Forest

- Variety
- TF_IDF
- 500 palabras



Género: mejor resultado

- 1000 palabras más frecuentes por género
- TF_IDF
- 700 palabras en el vocabulario
- Random Forest

Ajuste: 0.7279



Variedad: mejor resultado

- 1000 palabras más frecuentes por variedad
- TF_IDF
- 500 palabras en el vocabulario
- Random Forest

Ajuste: 0.9157

