PAN at CLEF 2022: Profiling Irony and Stereotype Spreaders on Twitter

Jose Arias Moncho

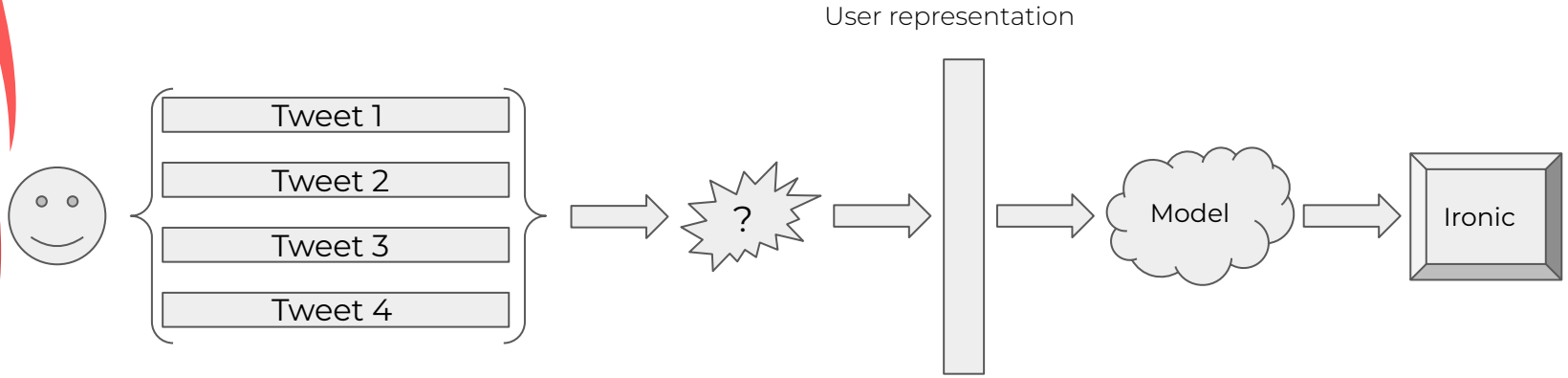
Victoria Beltrán Domínguez

Index

1. Introducción

Representamos a un usuario mediante un vector obtenido por:

1. Métodos estadísticos
2. BERT



2. Preproceso

Tokenizadores probados:

1. Borrar números y fechas

2. Anterior + Borrar #user#, #hashtag# and #url#

Además:

1. Pasar a minúscula el texto

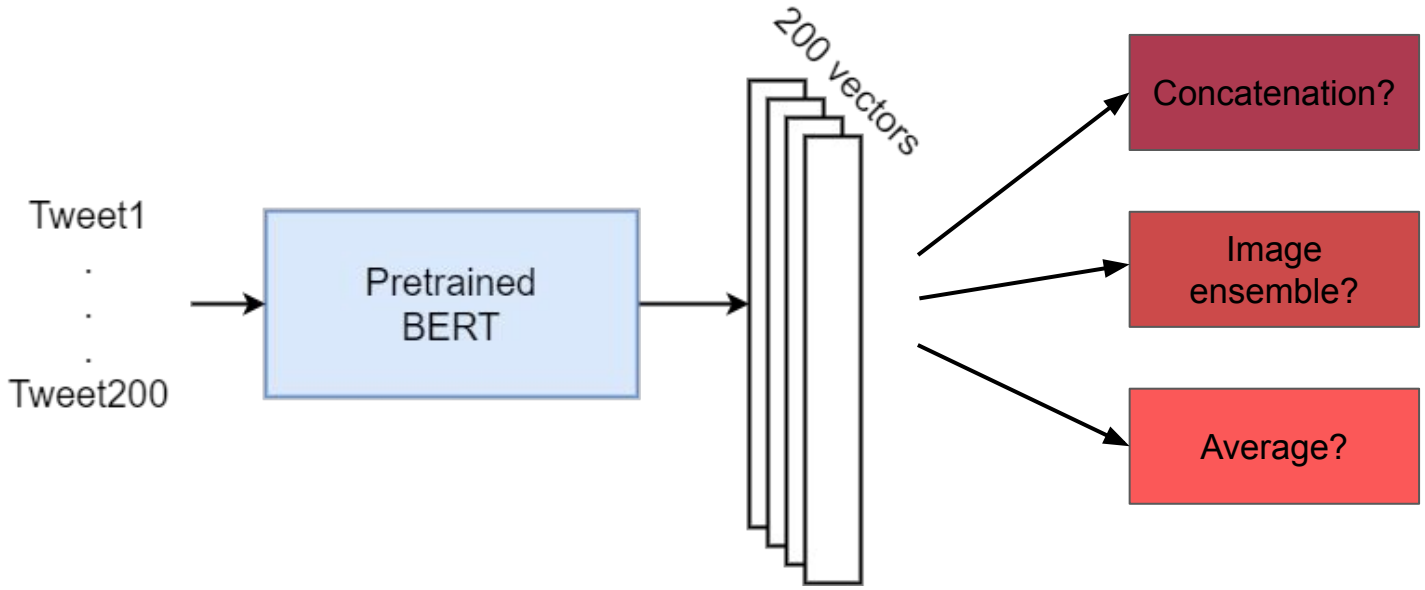
3. Vectorización Estadística

Vectorizers:

1. CountVectorizer
2. HashingVectorizer

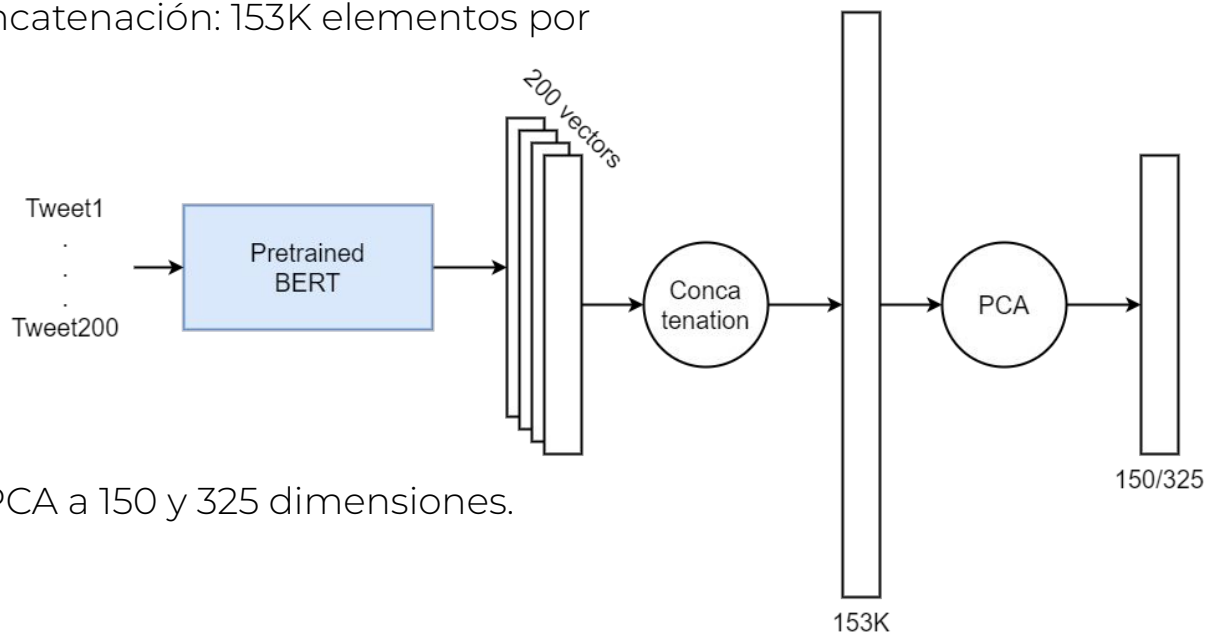
Convierten una colección de documentos de texto en una matriz de recuentos de tokens. La diferencia es que HashingVectorizer no almacena el vocabulario resultante, sino que relaciona cada token directamente a una columna de la matriz creada.

4. Vectorización basada en BERT



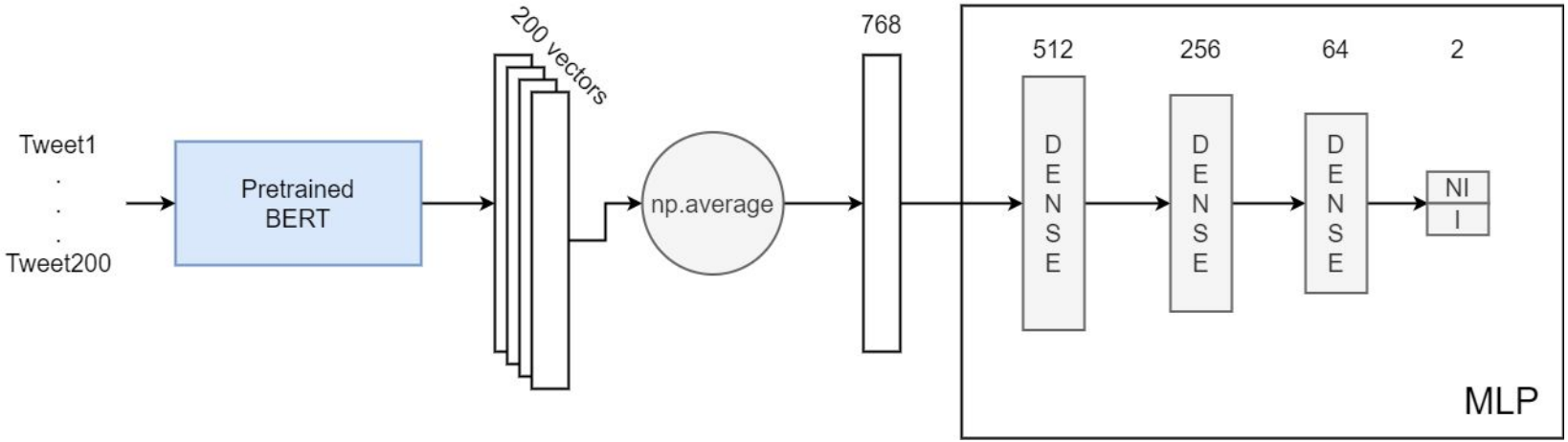
4. Vectorización basada en BERT: Concatenación

Problemas con la concatenación: 153K elementos por usuario

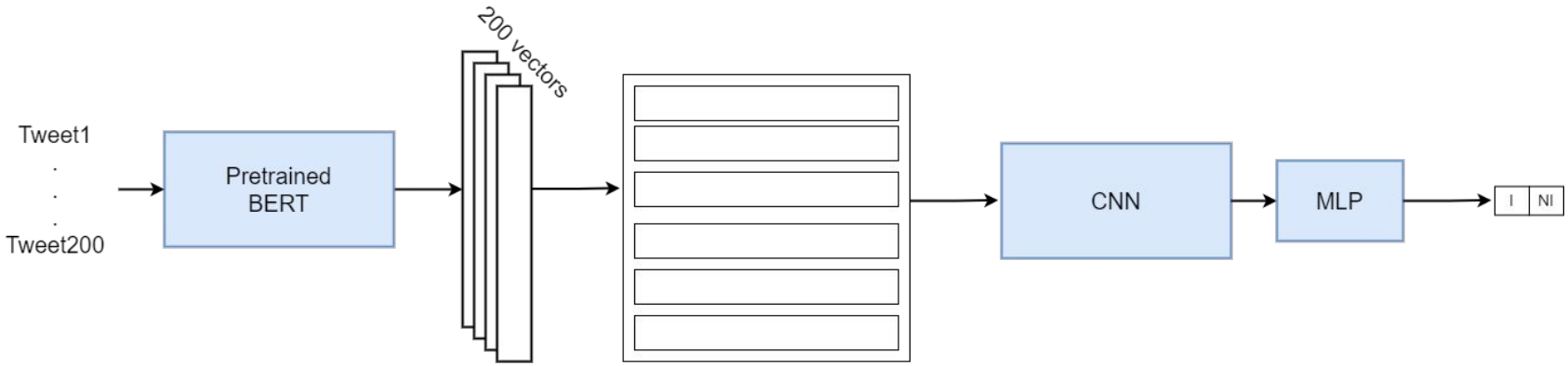


Soluciones? Aplicar PCA a 150 y 325 dimensiones.

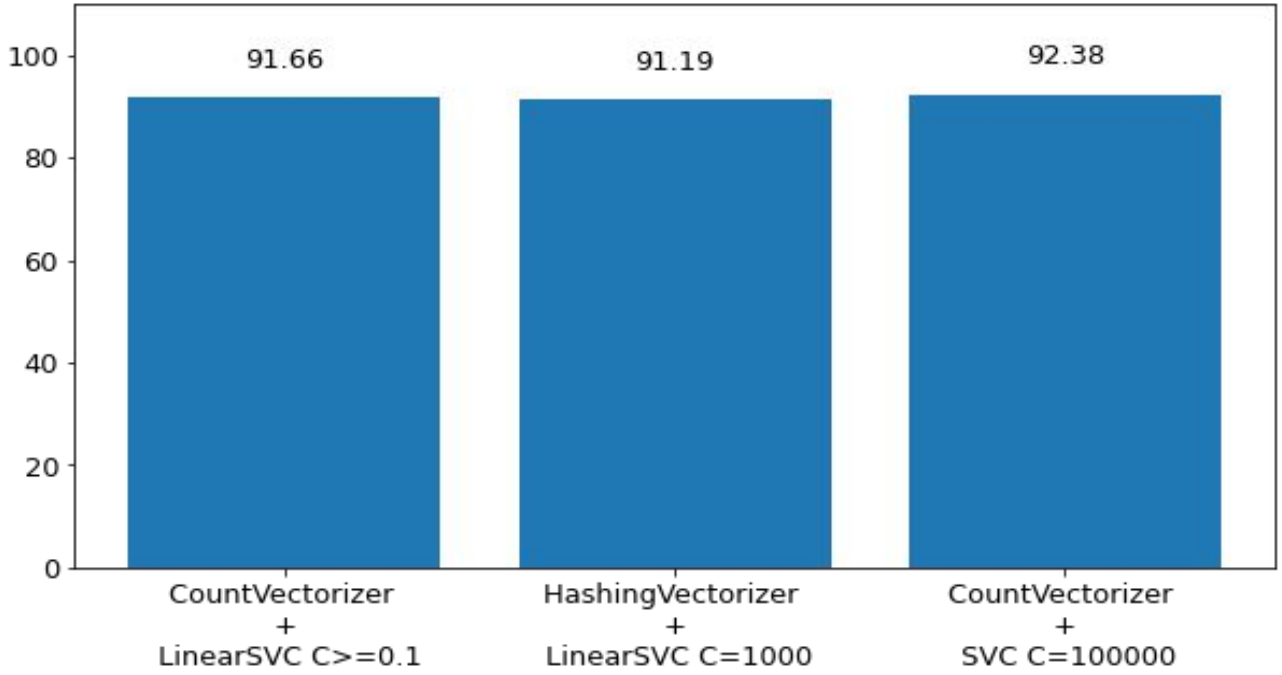
4. Vectorización basada en BERT: Media



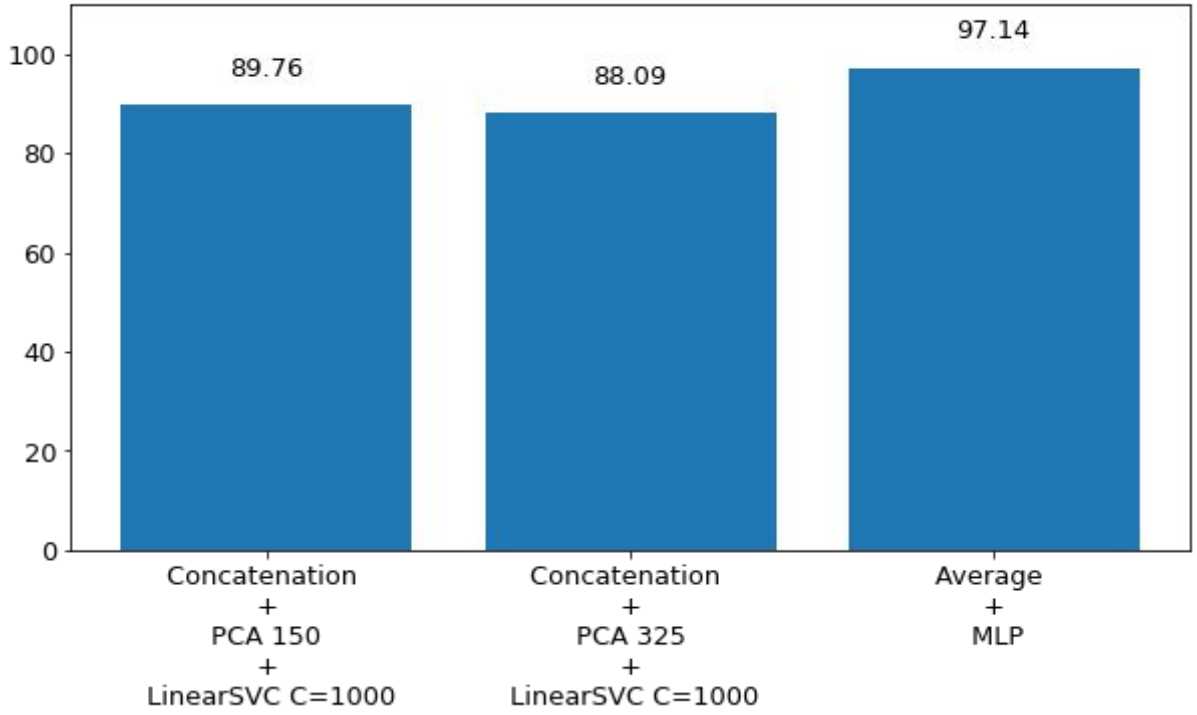
4. Vectorización basada en BERT: Image ensemble



5. Resultados: Vectorización Estadística



5. Resultados: Vectorizadores basados en BERT



5. Resultados: Otros resultados

Método	Precisión(%)
TfidfVectorizer + SVC (C=1000)	90.23
BertVectors CONCAT + PCA 325 + LinearSVC C=1000	88.09
BertVectors CONCAT + PCA 150 + MLP=[Dense(512) Dense(256) Dense(64)]	60.95
Image ensemble + CNN	50

6. Conclusión

Mejores resultados: Vectorización basada en BERT

¿Por qué?:

- Entrenado con millones de datos.
- Ayuda a resumir la información del lenguaje en cada tweet.
- Ayuda a discriminar las representaciones de los usuarios.
- Combinado con un MLP, consigue distinguir entre usuarios irónicos contra no irónicos.