PROFILING HATE SPEECH SPREADERS ON TWITTER

PAN SHARED TASK

APLICACIONES DE LA LINGÜÍSTICA COMPUTACIONAL

MIARFID, UPV

Jaime Ferrando Huertas, Javier Martínez Bernia

ÍNDICE

- Introducción
- Adquisición y preproceso de los datos
- Solución propuesta
- Experimentación y resultados
- Conclusiones

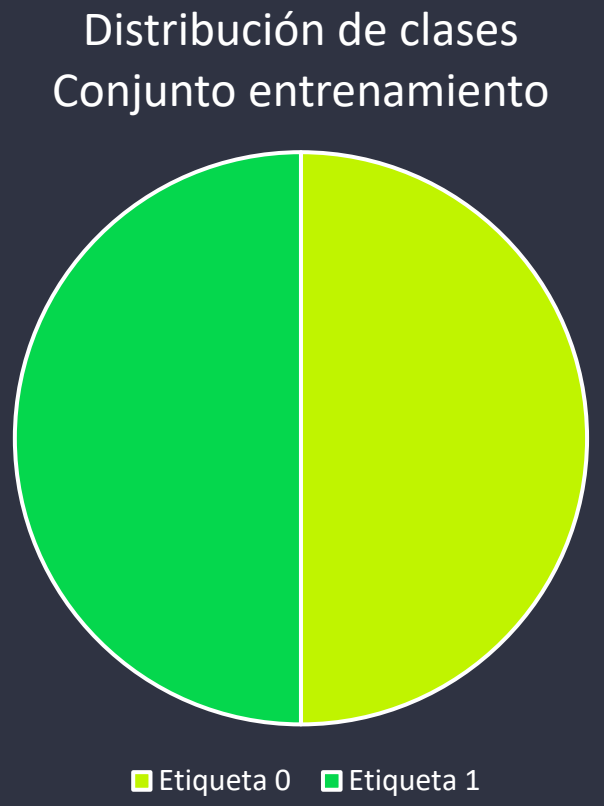
INTRODUCCIÓN

- PAN shared task
- Detectar si un autor difunde odio en twitter
- Español / Inglés

ADQUISICIÓN Y PREPROCESO DE LOS DATOS

ADQUISICIÓN DE LOS DATOS

- Conjunto de entrenamiento
  - 200 autores, 200 tweets por autor (ES/EN)
  - Etiquetas: 0, 1
- Parser
- Procesado en dos perspectivas
  - Tweets juntos
  - Tweet a tweet



PREPROCESO

- Borrado de etiquetas de twitter

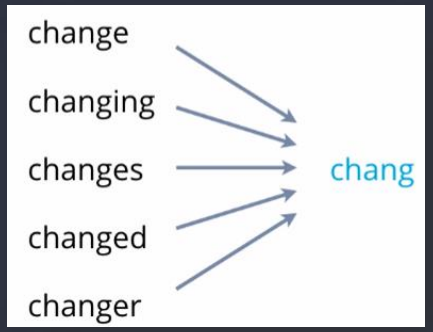
#HASHTAG#
#USER#
#RT#

- Sustitución de emojis

- Borrado de Stopwords

a, de, yo, mi, que…
a, about, and, the, it…

- Stemming



SOLUCIÓN PROPUESTA

SCIKIT-LEARN MODELS

1. Gradient Boosting Classifier
2. Linear Support Vector Machine
3. Non-Linear Support Vector Machine
4. Stochastic Gradient Descent Classifier
5. Multi-Layer Perceptron Classifier
6. K-Nearest Neighbors Classifier
7. Random Forest Classifier
8. Gaussian Naive-Bayes
9. Decision Tree Classifier

- Vectorizadores
  - Count Vectorizer
  - Tf-Idf Vectorizer

HUGGING FACE

- Twitter-Roberta-Base
- Modelo de Deep Learning basado en BERT
- Bi-directional Tranformer
- Preentrenado con datos en inglés

EXPERIMENTACIÓN Y RESULTADOS

EXPERIMENTACIÓN

Tweets juntos

- Validación cruzada en 10 bloques
- Modelos de Scikit-Learn
- Una muestra y una etiqueta por autor

Tweet a Tweet

- Modelo preentrenado de Hugging Face
- Validación con el conjunto de entrenamiento entero
- 200 muestras y 1 misma etiqueta
- Función para etiquetar las salidas

$$Label(n_{pos}, n_{neg}, ratio) = \begin{cases} 1 & \text{if } n_{neg} > ratio \cdot n_{pos} \\ 0 & \text{other cases} \end{cases}$$

TWEETS JUNTOS

Modelo	Vectorizador	Accuracy (es)	Accuracy (en)
Gradient Boosting Classifier	Count Vectorizer	0.75	0.69
Linear Support Vector Machine	Count Vectorizer	0.79	0.70
Non-Linear Support Vector Machine	Count Vectorizer	0.81	0.72
Stochastic Gradient Descent	Count Vectorizer	**0.83**	0.71
Multi-Layer Perceptron	Count Vectorizer	0.82	**0.75**
K-Nearest Neighbors	Count Vectorizer	0.76	0.62
Random Forest	Count Vectorizer	0.79	0.72
Gaussian Naive-Bayes	Count Vectorizer	0.75	0.61
Decision Tree	Count Vectorizer	0.68	0.61

TWEET A TWEET

Modelo	Preproceso	Ratio	Accuracy (en)
twitter-roberta-base	No	0.2	0.5
twitter-roberta-base	No	0.4	0.54
twitter-roberta-base	No	0.6	0.58
twitter-roberta-base	No	0.8	0.63
twitter-roberta-base	No	1.0	**0.66**
twitter-roberta-base	No	1.2	0.57
twitter-roberta-base	No	1.4	0.57
Multi-Layer Perceptron	Si	0.5	0.50

MODELOS SELECCIONADOS

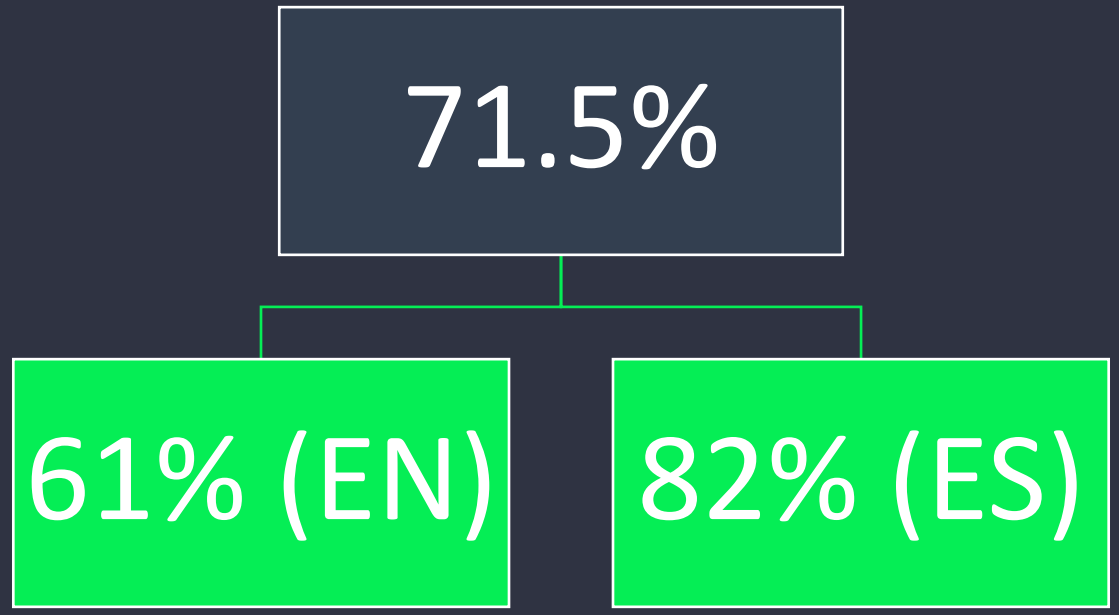
Español

- Stochastic Gradient Descent

Inglés

- Multi-Layer Perceptron

Accuracy en el conjunto de test



CONCLUSIONES

CONCLUSIONES

- Tarea Hate Speech
- Procesamiento del texto
- Buenos resultados en Español
- Modelo basado en Tranformer no acepta todos los tweets

GRACIAS POR SU ATENCIÓN

Jaime Ferrando Huertas
Javier Martínez Bernia