

# ¿Female or male?

# Text Mining en Social Media

Sergio Roca Martínez

Eusebio Soriano Benegas

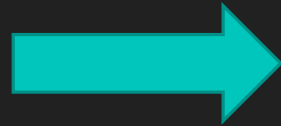
Jose Eduardo Reinoso Andrade

Carlos Peiró González

# Índice

- Introducción
- Propuesta
- Resultados
- Conclusiones

# Introducción



Identificar con la mayor precisión posible el género del autor de un tuit determinado.

# Dataset

## ○ DATASET

- Corpus training: 2800 ficheros .xml
- Corputs test: 1400 ficheros .xml

Cada .xml contiene 100 tuits

- fichero .txt → relaciona identificador tuit

- Género
- Variedad lingüística del español.

# Propuesta

Eliminamos signos de  
puntuación

Eliminamos tildes

**Bag of  
words**



```
graph LR; A[Bag of words] --> B[Eliminamos signos de puntuación]; A --> C[Eliminamos tildes]; A --> D[Stop words: determinantes...]; A --> E[Eliminamos palabras genéricas]; A --> F[Transformamos todo a minúsculas];
```

The diagram illustrates the preprocessing steps for a Bag of words model. A central teal square labeled 'Bag of words' has five arrows pointing outwards to different text blocks. Two arrows point left to 'Eliminamos signos de puntuación' and 'Eliminamos tildes'. Two arrows point right to 'Stop words: determinantes...' and 'Eliminamos palabras genéricas'. One arrow points down and to the right to 'Transformamos todo a minúsculas'.

Stop words:  
determinantes...

Eliminamos palabras  
genéricas

Transformamos todo a  
minúsculas

# Resultados

- **Baseline empleado**
  - Vocabulario: 100 palabras
  - Bag of words: 50 palabras

Modelo	Accuracy	kappa
SVM – sin eliminar términos	0.6529	0.3057
SVM – términos frecuentes por género	0.6736	0.3471

# Conclusiones

- Otras técnicas
  - Estudio de emoticonos.
  - Enlaces a fotos, webs...
  - Temas y tópicos tuits.
  - Contador numero de repeticiones por género.