¿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

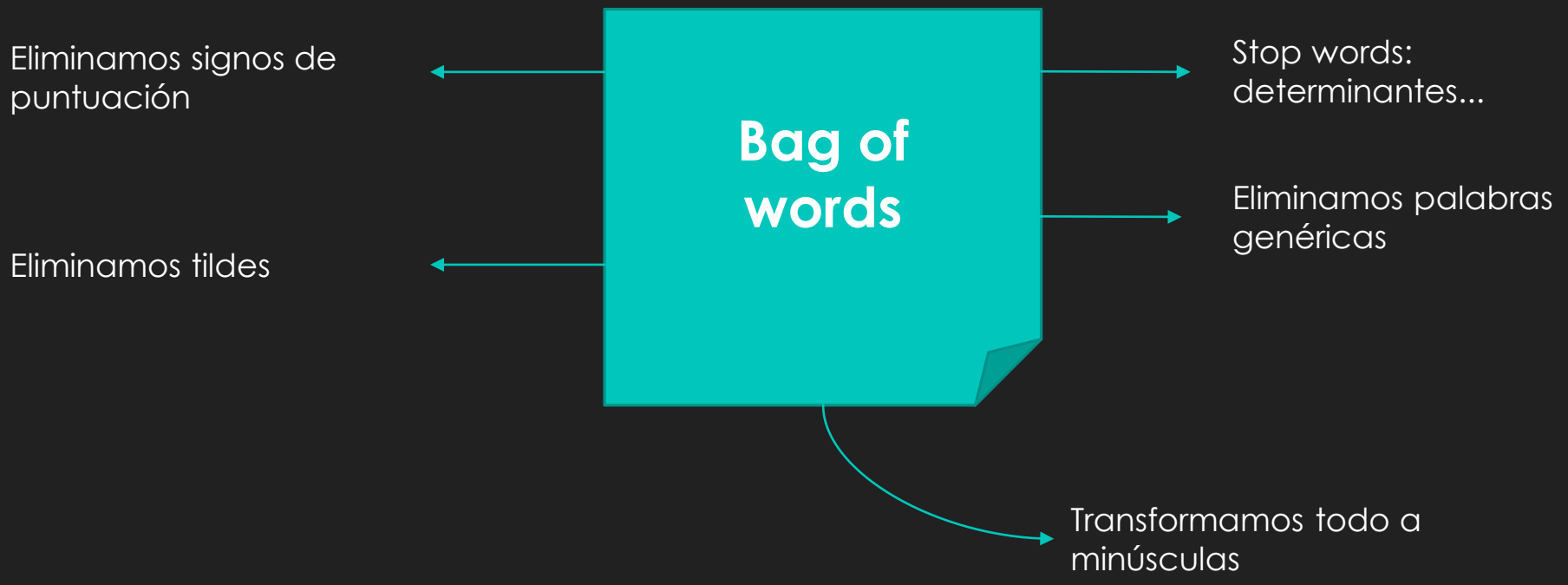
- **DATASET**
  - Corpus training: 2800 ficheros .xml
  - Corputs test: 1400 ficheros .xml

  } Cada .xml contine 100 tuits

  - fichero .txt → relaciona identificador tuit

    - Género
    - Variedad lingüística del español.

Propuesta


Eliminamos signos de puntuación
Eliminamos tildes
Bag of words
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 tópics tuits.
  - Contador numero de repeticiones por género.