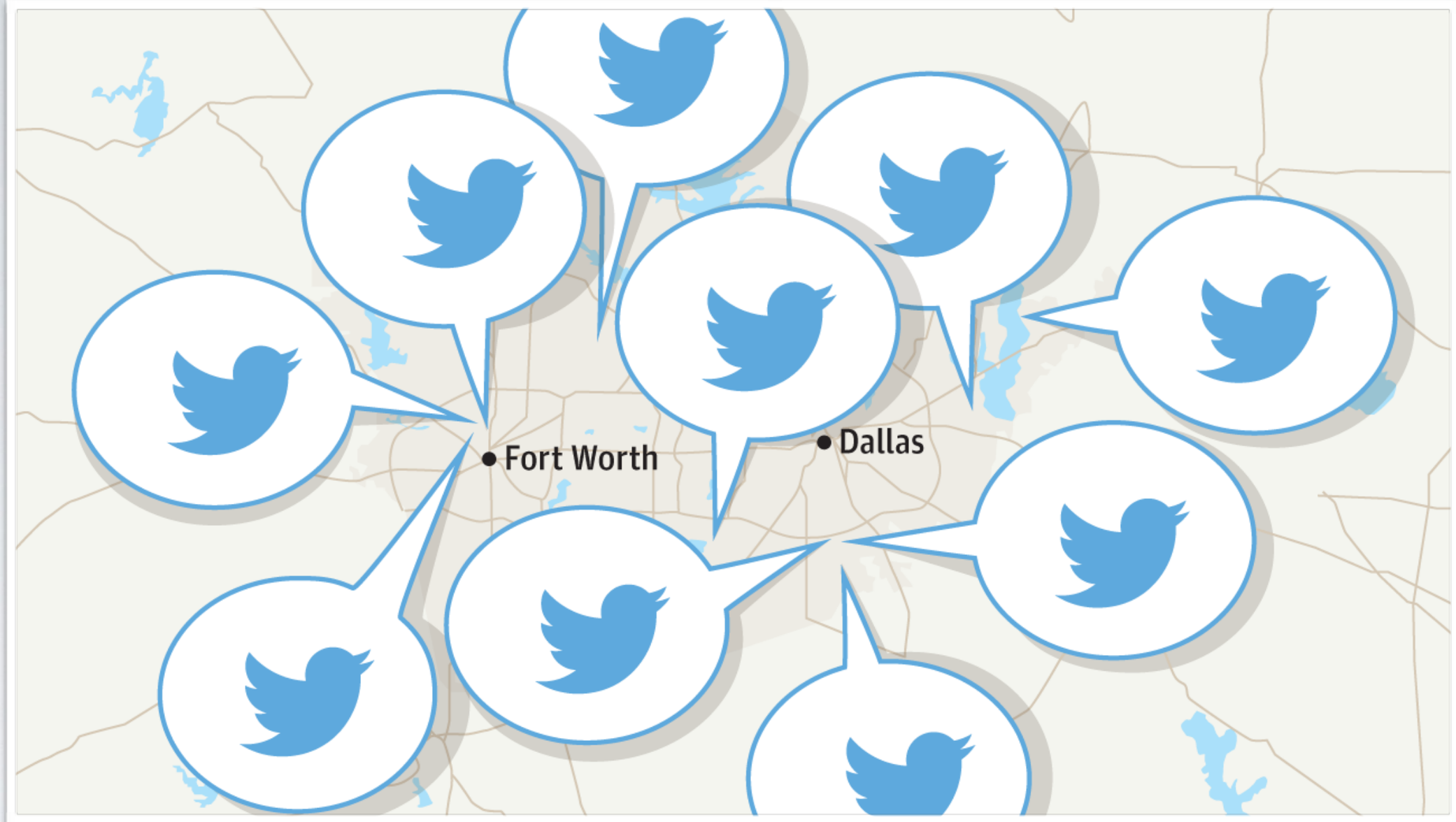
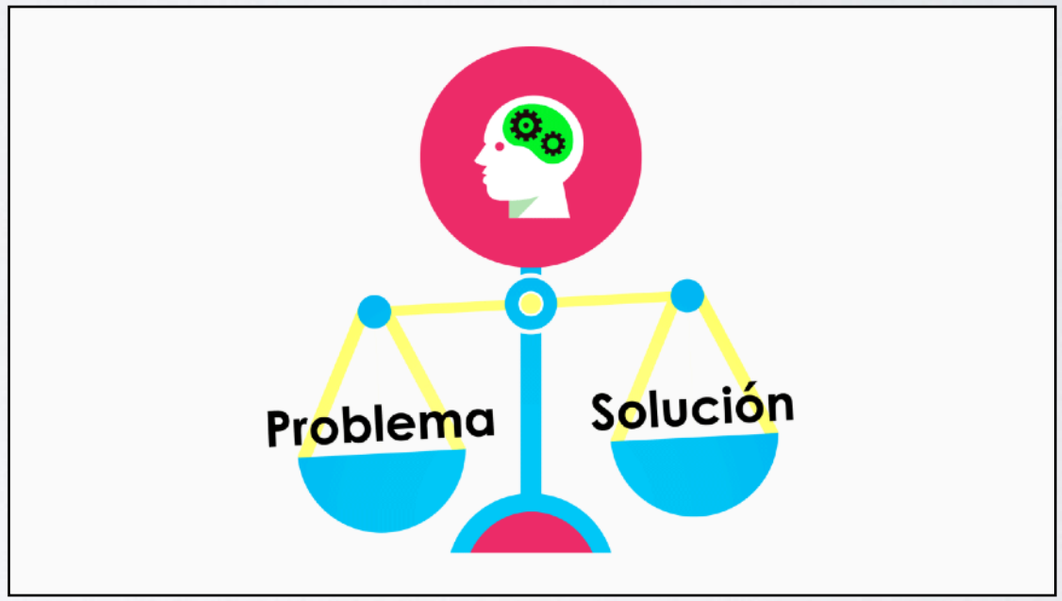


TEXT MINING

En social Media

PROBLEMA

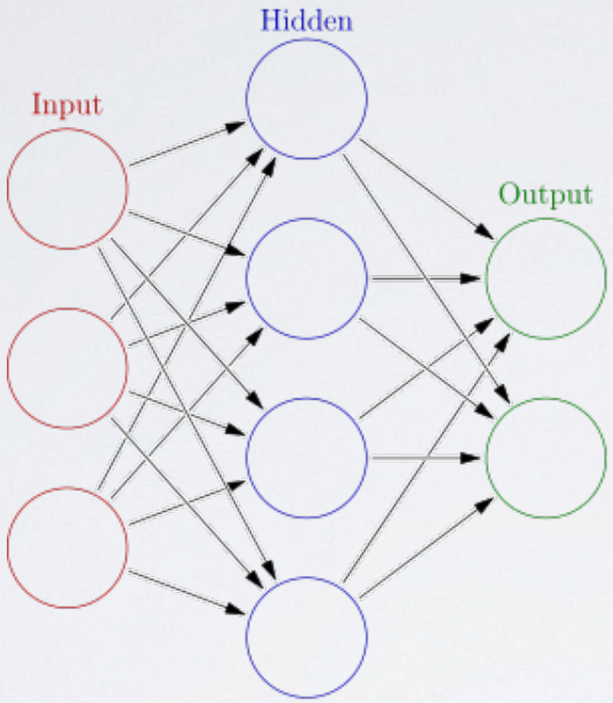
- Mejorar las predicciones del género y variedad
- Dataset de tweets: 300 autores por genero y lenguaje
- 100 tweets por autor
- Predicción a superar acc 0.66 en genero y 0.77 en variedad



LIMPIEZA DE DATOS

- Textos a minúsculas
- Eliminar signos de puntuación
- Eliminar números
- Stopwords: letras sueltas [a-z], preposiciones.
- Eliminar dos o más espacios en blanco.

MODELOS TESTEADOS



- SVM (lineal)
- Red neuronal
- Random Forest
- Naive Bayes

OTRAS PROPUESTAS

- ngrams (V-G)
- Mantener sólo la raíz de la palabra (V)
- Probar con las palabras más específicas de cada país (V)
- Última letra de la palabra(G)
- Longitud de las palabras (G)

SIGUIENTE!