# Minería de Texto
## Problemas SEGUNDO Parcial

## 1. Clasificación de Noticias Falsas

La **clasificación de noticias falsas** mediante el aprendizaje automático es de suma importancia en la era digital actual. El rápido crecimiento de las plataformas de redes sociales y los medios de comunicación online ha dado lugar a una abrumadora cantidad de información disponible para el público, lo que dificulta distinguir entre noticias reales y falsas. El aprendizaje automático, en particular las técnicas de procesamiento del lenguaje natural (PLN), desempeña un papel fundamental a la hora de abordar esta cuestión.

### Corpus:

- https://www.kaggle.com/datasets/sadmansakibmahi/fake-news-detection-dataset-with-pre-trained-model/data

### Notas metodológicas

- Técnicas que deben usarse para efectos comparativos:
    a) Un modelo de N-gramas (cualquier valor de N) y Word2Vec como modelos de representación base y un modelo de clasificación base– baseline–(SVM o random forest).
    b) Al menos un enfoque de representación alternativo (combinación de WE con N-gramas, o modelos tipo BERT)
    c) Al menos dos enfoques de clasificación para comparar contra el baseline (FFN, CNN, LSTM).

### Métricas:

1. **F1-Score (macro)**: Para evaluar el equilibrio entre precisión y exhaustividad, especialmente relevante ante posibles desbalances en la distribución de clases.

2. **Accuracy (Exactitud)**: Como medida general del rendimiento del clasificador.

3. **Matriz de Confusión**: Para analizar patrones de error específicos entre las clases ideológicas.

4. **Kappa de Cohen**: Para medir la concordancia entre las etiquetas predichas y las reales, corrigiendo el efecto del azar.

5. **AUC-ROC**: Para evaluar la capacidad discriminativa del modelo bajo diferentes umbrales de decisión.

## 2. Detección de similitud semántica

La **similitud semántica** es clave en el Procesamiento del Lenguaje Natural porque permite a los sistemas comprender el significado y la intención del lenguaje, más allá de la coincidencia de palabras. Gracias a ella, motores de búsqueda pueden ofrecer resultados más relevantes, los

sistemas de recomendación sugieren contenido por su sentido, y asistentes virtuales entienden la intención de los usuarios. También se aplica en resúmenes automáticos, clasificación y extracción de información, traducción más precisa, detección de plagio y análisis de sentimientos. Para medirla, se usan representaciones matemáticas del lenguaje (vectores), cuya cercanía refleja mayor similitud, aplicando técnicas como embeddings, la similitud coseno y modelos de transformadores como BERT o GPT. Se emplearán embeddings preentrenados (ej: Word2Vec, GloVe o FastText), combinados con estrategias de agregación vectorial (promedio, TF-IDF weighting) para representar oraciones. El modelo debe generar puntuaciones de similitud continuas que reflejen la cercanía conceptual entre textos, incluso cuando no compartan palabras clave literales.

## Corpus:

El primer conjunto de datos públicos de Quora está relacionado con el problema de identificar preguntas duplicadas. En Quora, un principio importante del producto es que debe haber una única página de preguntas para cada pregunta lógicamente distinta. Por ejemplo, las consultas «¿Cuál es el estado más poblado de EE. UU.?» y «¿Qué estado de Estados Unidos tiene más habitantes?» no deberían existir por separado en Quora, ya que la intención detrás de ambas es idéntica. Tener una página canónica para cada consulta lógicamente distinta hace que el intercambio de conocimientos sea más eficiente en muchos sentidos: por ejemplo, quienes buscan conocimientos pueden acceder a todas las respuestas a una pregunta en un solo lugar, y los escritores pueden llegar a un público más amplio que si ese público estuviera dividido entre varias páginas.

El conjunto de datos se basa en datos reales de Quora y ofrecerá a cualquiera la oportunidad de entrenar y probar modelos de equivalencia semántica.

- https://www.kaggle.com/datasets/quora/question-pairs-dataset

## Notas metodológicas

- Técnicas que deben usarse para efectos comparativos:
    a) Un modelo LSA (TF-IDF+SVD) o Word2Vec como modelo de representación base (baseline).
    b) Al menos un enfoque de representación alternativo (sentence-BERT, paraphrase-multilingual-mpnet-base-v2, fine-tuned BERT)
    c) Al menos dos medidas de similitud adicionales a la distancia euclidiana y distancia coseno (similitud de Dice, divergencia Kullback-Leibler, distancia de Mahalanobis) para determinar la salida del modelo.

## Métricas:

1. **Coeficiente de correlación de Pearson (r)**: Para medir la correlación lineal entre las puntuaciones de similitud predichas y las anotaciones de referencia.

2. **Coeficiente de correlación de Spearman (ρ)**: Para evaluar la correlación monótona entre los rangos de similitud predichos y los reales (robusto ante relaciones no lineales).

3. **Error Cuadrático Medio (MSE)**: Para cuantificar la magnitud promedio de los errores al cuadrado en las predicciones.

4. **Error Absoluto Medio (MAE)**: Para medir la desviación promedio absoluta entre puntuaciones predichas y reales.

5. **Precisión @ Top-K (ej: P@5)**: Para evaluar la capacidad del modelo en tareas de recuperación (ej: identificar los 5 pares más similares en un conjunto).

## 3. Análisis de sentimientos 1 y 2

El **análisis de sentimientos** es una herramienta del Procesamiento del Lenguaje Natural que permite comprender las emociones y opiniones expresadas en textos, y tiene un gran valor en ámbitos sociales. Puede usarse para detectar el estado de ánimo de comunidades en redes sociales tras una crisis, identificar señales tempranas de depresión o ansiedad en foros de apoyo, o analizar la percepción ciudadana sobre políticas públicas. También es útil para organizaciones sociales y educativas, ya que ayuda a monitorear el bienestar emocional de estudiantes o comunidades vulnerables. Al captar no solo si un mensaje es positivo o negativo, sino también matices como la ironía o la frustración, el análisis de sentimientos contribuye a una mejor toma de decisiones y a diseñar intervenciones más humanas y efectivas.

El objetivo es desarrollar un sistema de clasificación multiclase que asigne automáticamente categorías emocionales (ej: miedo, tristeza, ira, esperanza) a mensajes breves (tweets), combinando diccionarios lexicales predefinidos con técnicas de aprendizaje automático. El desafío incluye manejar el lenguaje informal, ambigüedad contextual y el desbalance natural entre emociones (ej: predominio de "tristeza" sobre "esperanza"). Se permitirá el uso de cualquier técnica de representación textual (TF-IDF, embeddings, etc.), pero se requerirá la integración de un diccionario léxico de emociones como característica adicional o para validación de resultados.

### Corpus

1. Tweets en español: https://www.kaggle.com/datasets/philipsanm/sentiment-analysis-in-spanish-tweets
2. Reseñas de libros: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

### Diccionarios léxicos

- ML-Senticon: https://github.com/ITALIC-US/ML-Senticon
- VADER: https://github.com/cjhutto/vaderSentiment
- EmoLex: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

### Notas metodológicas

- El diccionario puede usarse como:
    a) Características adicionales (ej: conteo de palabras por emoción + TF-IDF).
    b) Validación (comparar predicciones del modelo con categorías del diccionario).
- Técnicas que deben usarse para efectos comparativos:

a) Un modelo de N-gramas (cualquier valor de N) o Word2Vec como modelo de representación base (baseline) y un modelo de clasificación base (SVM o random forest).
b) Al menos dos enfoques de representación alternativos (modelos tipo BERT)
c) Al menos un enfoque de clasificación para comparar contra el baseline (redes residuales, CNN, LSTM).

- Otros modelos comparativos: https://github.com/sentiment-analysis-spanish/sentiment-spanish

### Métricas:

1. **F1-Score (macro)**: Para evaluar el balance entre precisión y exhaustividad en todas las clases emocionales, crítico ante el desbalance natural.

2. **Matriz de Confusión**: Para identificar errores específicos entre emociones (ej: confusión entre "ira" y "miedo").

3. **Accuracy (Exactitud)**: Como referencia general del rendimiento.

4. **Kappa de Cohen**: Para medir concordancia inter-anotadores entre predicciones y etiquetas reales.

5. **Precisión por Clase**: Para detectar emociones con bajo rendimiento (ej: "sorpresa" vs "alegría").

## 4. Sistemas de recomendación

Los **sistemas de recomendación** son herramientas de inteligencia artificial que ayudan a filtrar y priorizar información para que los usuarios encuentren de manera más rápida lo que les interesa. Funcionan analizando datos sobre preferencias, comportamientos y similitudes entre usuarios o elementos. Más allá del comercio electrónico, estos sistemas tienen un fuerte impacto social y educativo: pueden recomendar recursos de aprendizaje personalizados, contenidos culturales relevantes, o incluso información crítica en contextos de salud pública. Su propósito central es reducir la sobrecarga de información y ofrecer sugerencias útiles y significativas para cada persona.

### Corpus:

En el caso de Amazon, el **corpus de reseñas de libros** es una de las fuentes más usadas para investigar y entrenar sistemas de recomendación y modelos de Procesamiento del Lenguaje Natural. Este corpus está compuesto por millones de reseñas de usuarios que incluyen tanto texto libre (opiniones, comentarios, críticas) como metadatos (calificaciones con estrellas, fecha, autor de la reseña, e información sobre el libro). Su riqueza radica en que combina valoraciones explícitas (las estrellas) con valoraciones implícitas (el análisis del lenguaje en las reseñas), lo que permite estudiar patrones de preferencia, tendencias de lectura y relaciones entre libros. Gracias a este tipo de datos, los investigadores pueden entrenar modelos de recomendación más precisos, que no solo sugieren títulos basados en compras previas, sino también en afinidad temática, estilo narrativo o similitud semántica entre reseñas.

- Reseñas de libros: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

## Notas metodológicas

- Técnicas que deben usarse para efectos comparativos:
    a) Un modelo de N-gramas (cualquier valor de N) como baseline. Para usar N-gramas en un sistema de recomendación de libros, se trata el historial de lectura de cada usuario como una secuencia ordenada de libros (o sus identificadores), y se generan subsecuencias de N elementos (por ejemplo, bigramas o trigramas). Luego, se calcula la frecuencia con que un libro sigue a una determinada secuencia en los datos de todos los usuarios, lo que permite estimar la probabilidad de que un libro sea el siguiente en la secuencia actual de un usuario. Por ejemplo, si muchos usuarios leen *Rayuela* seguido de *Ficciones* y luego *El Aleph*, el sistema recomendará *El Aleph* a quien haya leído los dos primeros. Este enfoque captura patrones secuenciales sencillos y es especialmente útil cuando se combinan libros por autor, género o tema para evitar datos demasiado dispersos.
    b) Un enfoque de representación único o combinado (N-gramas, TF-IDF o WE)
    c) Al menos un enfoque alternativo para comparar contra el baseline (factorización matricial, redes neuronales, modelos basados en grafos o aprendizaje de representaciones: Item2Vec, Graph Neural Networks (GNNs), LightGCN).

## Métricas:

1. **Error Absoluto Medio (MAE)**: Mide la desviación promedio de las calificaciones predichas respecto a las reales. Se utiliza cuando se predicen calificaciones explícitas.

2. **Error Cuadrático Medio (RMSE)**: Similar al MAE, pero penaliza más los errores grandes, lo que puede ser útil para detectar las predicciones más erróneas.

3. **Precision@K**: Mide la proporción de elementos recomendados en los K primeros resultados que son realmente relevantes para el usuario.

4. **Recall@K**: Mide la proporción de todos los elementos relevantes que se incluyen en las K primeras recomendaciones.

5. **MAP (Mean Average Precision)**: Es una medida de la precisión de un sistema de clasificación que considera la cantidad de elementos relevantes y su posición en la lista.