Sistema de Procesamiento de Noticias en Español

Equipo 6:

- Andrés Alexis Galvis Herrera
- Juan Esteban Mejía Espejo
- Juan José Zapata Cadavid
- Maria Camila Zapata Arrubla

Introducción y Descripción del Dataset

El proyecto utiliza el conjunto de datos "Spanish News Classification" de Kaggle, que contiene aproximadamente 1200 noticias en español clasificadas en 7 categorías: Macroeconomía (28.0%), Alianzas (20.3%), Innovación (16.1%), Regulaciones (11.7%), Sostenibilidad (11.3%), Otra (10.7%) y Reputación (2.1%). Esta distribución proporciona una base adecuada para desarrollar y evaluar los componentes del sistema de procesamiento de lenguaje natural.

1. Clasificación Automática de Noticias

Implementación

Se desarrolló un sistema de clasificación automática siguiendo estos pasos:

1. **Preprocesamiento**: Eliminación de columnas irrelevantes, limpieza de texto conservando caracteres especiales del español, conversión a minúsculas y normalización de espacios.

2. **Arquitectura**: Se utilizó BERT multilingüe (bert-base-multilingual-cased) con los siguientes parámetros:

    - Longitud máxima: 128 tokens
    - Tamaño de lote: 16
    - Tasa de aprendizaje: 2e-5
    - Épocas: 10

3. **Gestión de datos**: División estratificada (90% entrenamiento, 10% validación) para mantener la distribución de clases.

4. **Optimización**: Optimizador AdamW con programación lineal de tasa de aprendizaje y gradient clipping.

Resultados

El modelo alcanzó:

- Accuracy: 84.43%

- F1-Score: 0.8433
- Recall: 0.8443
- Precision: 0.8509

El modelo mostró excelente desempeño en categorías mayoritarias como "Macroeconomía" y "Alianzas", aunque presenta un rendimiento variable en categorías minoritarias y cierta confusión entre categorías semánticamente similares.

2. Resumen Abstractivo de Noticias

Implementación

1. **Modelo**: Se seleccionó "mrm8488/bert2bert_shared-spanish-finetuned-summarization", basado en BETO y entrenado específicamente para resúmenes en español.

2. **Preprocesamiento**: Se implementó un procedimiento de truncamiento para limitar la longitud del texto a 510 tokens, respetando los límites del modelo.

3. **Generación**: Se configuró el pipeline con longitud máxima de 130 caracteres y mínima de 30 caracteres.

4. **Evaluación**: Se utilizaron métricas ROUGE (Recall-Oriented Understudy for Gisting Evaluation) para evaluar los resúmenes generados:

    - ROUGE-1: Coincidencia de unigramas
    - ROUGE-2: Coincidencia de bigramas
    - ROUGE-L: Coincidencia de la subsecuencia común más larga

Resultados

Las métricas obtenidas fueron:

- ROUGE-1: 0.1498
- ROUGE-2: 0.1309
- ROUGE-L: 0.1384

Estos valores relativamente bajos indican una coincidencia limitada entre los textos originales y los resúmenes. Posibles causas incluyen el truncamiento de textos, la complejidad lingüística de las noticias y limitaciones del modelo para el dominio específico.

3. Identificación de Temas con LDA

Implementación

1. **Preprocesamiento especializado**:

    - Tokenización con NLTK
    - Eliminación de stopwords en español

- Lematización con spaCy
- Filtrado por longitud (>2 caracteres)
- Filtrado por frecuencia (percentiles 5%-95%)

2. **Modelado**:

   - Evaluación de modelos LDA con diferentes números de temas (4-7)
   - Análisis de coherencia (c_v y u_mass)
   - Implementación final con 7 temas
   - 10 pases para asegurar convergencia y semilla 42 para reproducibilidad

3. **Visualización y análisis**:

   - Heatmaps de distribución de temas por documento
   - Análisis de asignación de documentos
   - Comparación con categorías originales

Resultados

El modelo con 7 temas obtuvo la mejor coherencia (c_v: 0.562), identificando estas agrupaciones temáticas:

- **Tema 0**: Industria y economía regional (174 documentos)
- **Tema 1**: Transporte y movilidad (175 documentos)
- **Tema 2**: Banca y finanzas (222 documentos)
- **Tema 3**: Tecnología y startups (168 documentos)
- **Tema 4**: Temas sociales y medioambientales (159 documentos)
- **Tema 5**: Energía y recursos (148 documentos)
- **Tema 6**: Tecnología avanzada y seguridad (171 documentos)

Se observó una clara correlación entre los temas descubiertos automáticamente y las categorías originales, con una distribución balanceada que sugiere una estructura temática bien diferenciada.

4. Conclusiones Generales

Desempeño Comparativo

Componente	Rendimiento	Observaciones
Clasificación	Excelente (>84%)	Alta precisión, útil para aplicaciones que requieren confiabilidad
Resumen	Moderado (13-15%)	Tarea más desafiante, requiere modelos más especializados
Modelado de temas	Bueno (coherencia 0.562)	Descubrió estructura similar a la clasificación manual

Hallazgos Clave

1. **Efectividad por Tarea**:

- Los modelos multilingües como BERT son efectivos para clasificación en español
- El modelado de temas con LDA es robusto con preprocesamiento adecuado
- El resumen abstractivo en español presenta oportunidades de mejora

2. **Integración de Componentes**:

   - El clasificador proporciona categorías predefinidas
   - El modelador de temas descubre estructuras emergentes
   - El resumidor condensa el contenido para análisis rápido

Recomendaciones para Trabajo Futuro

1. **Mejora del Componente de Resumen**:

   - Experimentar con modelos más recientes o específicos para español
   - Implementar fine-tuning adicional con corpus de noticias
   - Explorar evaluación humana complementaria a ROUGE

2. **Refinamiento del Modelado de Temas**:

   - Implementar modelado dinámico para cambios temporales
   - Explorar asignaciones múltiples de temas por documento
   - Integrar embeddings contextuales de BERT

3. **Extensión del Sistema**:

   - Desarrollar análisis de sentimiento para noticias económicas
   - Implementar detección de eventos y tendencias
   - Crear visualizaciones interactivas de evolución temática

El sistema establece una base sólida para el procesamiento automático de noticias en español, con resultados particularmente prometedores en clasificación y modelado de temas.