Tecnicatura en Ciencia de Datos e Inteligencia Artificial

Técnicas del Procesamiento del Habla

Informe Final PNL

EQUIPO DOCENTE

Mainero, Alejandro

GOOGLE COLAB

Enlace

GRUPO DATAMINDS

1. **Ayán, Trinidad**
2. **Giordano, Ariel**
3. **Herrera, Edgar Fabián**
4. **Quiroga, Fernanda**
5. **Siccardi, Luis**

CICLO LECTIVO 2025

Informe Final – Procesamiento del Habla Basado en PLN

a. Resumen del proyecto

Se diseñó y evaluó un **clasificador de sentimiento** para reseñas en español del *IMDB Dataset of 50K Movie Reviews*.

El flujo de trabajo fue:

1. **Ingesta y exploración**

   `pandas` para carga y análisis exploratorio.

   Revisión de columnas `review_es` y `sentimiento`.

2. **Limpieza de texto**

   Expresiones regulares (`re.sub`) para eliminar signos de puntuación y caracteres especiales.

   Conversión a minúsculas y normalización opcional de acentos con `unidecode`.

3. **Normalización léxica**

   Stemming: `SnowballStemmer` (ES).

   Lematización: `spaCy es_core_news_sm`.

   Stop-Words: filtro incorporado en `CountVectorizer(stop_words='spanish')`.

4. **Vectorización**

   Modelo **Bolsa de Palabras (BoW)** mediante `CountVectorizer`, una elección deliberada por su bajo costo computacional y transparencia para la etapa didáctica.

5. **Entrenamiento**

   Clasificador **Multinomial Naïve Bayes** – robusto para BoW y escalable a vocabularios grandes.

6. **Evaluación**

   Métricas: *accuracy, precision, recall, F1-score* y **tamaño de vocabulario** para siete variantes (baseline, tres técnicas individuales y tres combinaciones).

El objetivo es **cuantificar cómo las técnicas de preprocesamiento influyen simultáneamente en la precisión y en la comprensión del vocabulario** del modelo.

b. Análisis de resultados

Variante	Precisión	F1-score	Vocabulario
Baseline	74 %	75.2 %	15 432
Stemming	76.2 %	77.1 %	12 526
Lemmatización	75.5 %	76.2 %	12 233
Stop-Words	72.5 %	73.2 %	9 298
Stemming + Stop-Words	**79 %**	**79.5 %**	8 147
Lemmat. + Stop-Words	77.8 %	78.4 %	**7 856**

Mayor aumento de precisión: la combinación **Stemming + Stop-Words** elevó la precisión 5% sobre el baseline (de 74 % a 79 %).

Máxima reducción de vocabulario: **Lematización + Stop-Words** comprimió el léxico 49 % (de 15 432 a 7 856 términos).

¿Reducción ⇄ rendimiento?
Una poda *moderada* (Stemming o Lematización) mejoró todas las métricas.
Una poda *agresiva* sin normalización (solo Stop-Words) degradó la precisión pese a reducir 40 % el vocabulario.
Combinar **normalización morfológica + filtrado de Stop-Words** recompone la pérdida semántica y ofrece el mejor *trade-off* entre compresión y desempeño.

Técnica más eficiente: **Stemming + Stop-Words** — mantiene un léxico manejable (-47 %) y lidera en precisión/F1, ideal para modelos ligeros destinados a producción en tiempo real.

c. Reflexión crítica

Impacto del preprocesamiento

Una cadena de limpieza bien calibrada **reduce el ruido, homogeneiza variantes morfológicas y mejora la densidad de información** de los vectores. Este caso muestra ganancias de hasta +5% de precisión con un vocabulario a la mitad del original, validando la hipótesis de que *"menos, pero mejor"* favorece a clasificadores probabilísticos.

Limitaciones encontradas

- **Modelo BoW** ignora orden y contexto; las secuencias “no me gusta” y “me gusta no” son indistinguibles.
- **Naïve Bayes** asume independencia de términos, simplificación que no captura dependencias semánticas.
- **Coste en procesamiento**: lematizar con spaCy incrementa el tiempo ~4 × en relación con un stemmer rule-based.
- **Dependencia lingüística**: Snowball y spaCy están entrenados en español estándar; coloquialismos o jerga local pueden escapar al lematizador.

Mejoras futuras

- Sustituir BoW por **TF-IDF** o **word-embeddings** (e.g. FastText-ES) para incorporar peso contextual.
- Experimentar con **n-gramas** y modelos **SVM o Redes Neuronales** sobre las mismas features para evaluar si el preprocesamiento óptimo se mantiene.
- Aplicar **validación cruzada estratificada** y *hyper-parameter tuning* con *GridSearch* para afinar α de Laplace y umbrales de frecuencia.
- Incorporar **data-augmentation** sintáctica (paráfrasis neuronales) para robustecer el modelo ante sinónimos y variaciones dialectales.

Aplicabilidad a otros contextos

- Sistemas de **análisis de reputación** en e-commerce y métricas NPS.
- **Monitoreo de redes sociales** para detección temprana de crisis de marca.
- **Soporte al cliente**: enrutado de tickets por sentimiento/urgencia.
- **Educación**: evaluación automática de retroalimentación estudiantil.

La arquitectura y la bitácora empleadas en el *notebook* proporcionan un **patrón reproducible** para cualquier proyecto de PLN donde el balance entre *precisión* ↔ *ligereza* sea crítico (p.ej. despliegue en dispositivos edge o APIs de baja latencia).

Conclusión breve

La investigación confirma que **la combinación de normalización morfológica con filtrado de Stop-Words maximiza la eficacia** para clasificadores basados en BoW y Naïve Bayes. No obstante, la naturaleza bag-of-words seguirá limitando la comprensión semántica; evolucionar hacia representaciones contextuales será la vía para superar los techos detectados y trasladar los aprendizajes a dominios más complejos.