



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



FACULTAD DE
INGENIERÍA

FronteiraNLP: modelando el contacto español-portugués en Uruguay

Informe de Trabajo Final de Licenciatura presentado por

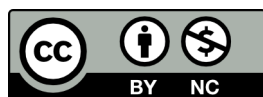
Rodrigo Souza

en cumplimiento parcial de los requerimientos para la graduación de la carrera
de Licenciatura en Computación de Facultad de Ingeniería de la Universidad
de la República

Supervisores

Luis Chiruzzo
Santiago Góngora

Montevideo, 7 de julio de 2026



FronteiraNLP: modelando el contacto español-portugués en Uruguay por Rodrigo Souza tiene licencia [CC Atribución - No Comercial 4.0](#).

Agradecimientos

Quiero expresar mi más sincero agradecimiento a mis tutores, por su compromiso, orientación y acompañamiento constante durante el desarrollo de esta tesis. Sus conocimientos, sugerencias y tiempo dedicado fueron fundamentales en cada etapa del trabajo y contribuyeron significativamente a su resultado final.

Agradezco especialmente a Bárbara Garrido por facilitar el acceso al corpus utilizado en esta investigación. Su colaboración y disposición hicieron posible el desarrollo de este trabajo, ya que sin esos datos esta investigación no habría podido llevarse a cabo.

Finalmente, quiero agradecer a mi familia y a mis amigos por su apoyo incondicional a lo largo de este proceso y de toda mi formación. Su paciencia, comprensión y aliento constante fueron un pilar fundamental para alcanzar esta meta.

Miña lingua le saca la lingua al dicionario
baila uma cumbia insima dus mapa
i fas com a túnica i a moña
uma cometa
pra voar
livre i solta pelu seu.

Des

Fabián Severo
NOITE NU NORTE

Resumen

Esta tesis presenta una serie de esfuerzos iniciales para aplicar técnicas de procesamiento de lenguaje natural (PLN) al estudio del portuñol, una variedad lingüística que surge del contacto sostenido entre el español y el portugués en la frontera Uruguay-Brasil. El objetivo principal fue desarrollar y evaluar métodos capaces de clasificar automáticamente cada *token* en una de seis categorías: español (*es*), portugués (*pt*), híbrido (*mix*), entidad nombrada (*ne*), extranjero (*foreign*) y otros símbolos o marcadores (*other*).

Como base experimental, se anotó manualmente un corpus previamente existente, compuesto por transcripciones orales de hablantes de la región norte de Uruguay. Este corpus fue preprocesado, tokenizado y dividido en conjuntos de entrenamiento, desarrollo y prueba, permitiendo realizar experimentos controlados bajo un marco reproducible de evaluación.

Se exploraron tres enfoques principales. En primer lugar, se implementaron heurísticas basadas en reglas y diccionarios de español, portugués e inglés. Este método, aunque simple e interpretable, alcanzó una *accuracy* de 0,88 gracias a su buen desempeño en las clases mayoritarias, pero mostró limitaciones notorias en las categorías minoritarias, especialmente en los tokens híbridos (*mix*).

En segundo lugar, se aplicó un modelo estadístico *Naive Bayes* en su variante *Complement*, utilizando representaciones de *n*-gramas de caracteres. Este enfoque logró mejorar parcialmente el reconocimiento de tokens híbridos ($F1=0,54$ frente a 0,40 de las heurísticas), aunque su desempeño global no superó a las reglas, con *accuracy*=0,85 y *Macro-F1*=0,64.

Por último, se evaluaron modelos neuronales basados en **BERT**. Se llevó a cabo un *fine-tuning* de dos variantes: el modelo multilingüe **bert-base-multilingual-cased** y **BERTinho**, entrenado en gallego, una lengua en contacto histórico con el español y el portugués. Ambos modelos lograron el mejor equilibrio entre clases: *mBERT* obtuvo *Macro-F1*=0,67 y *BERTinho* 0,66 en el conjunto de prueba, destacándose especialmente en la detección de tokens híbridos (*mix*).

Los resultados confirman que, incluso en este contexto de escasez de datos y fuerte variación lingüística, los enfoques basados en representaciones contextuales profundas ofrecen mayor robustez y capacidad de generalización frente a los métodos estadísticos y heurísticos. El estudio demuestra que el portuñol puede ser modelado computacionalmente mediante técnicas modernas de PLN y sienta las bases para el desarrollo de recursos y modelos específicos que contribuyan al análisis y preservación de las variedades lingüísticas en contacto.

Palabras clave: Portugués, Procesamiento de Lenguaje Natural, Clasificación de tokens, Naive Bayes, BERT, Heurísticas

Índice general

1. Introducción	1
2. Marco Teórico	4
2.1. El portugués uruguayo y el portuñol	4
2.2. Alternancia de códigos (Code-switching)	6
2.2.1. Tokens híbridos (<i>mix</i>) y su complejidad para el PLN . . .	7
2.3. Lenguas de escasos recursos en PLN	7
2.4. Corpus existentes para el portuñol	8
2.5. Detección automática de idioma en PLN	10
2.5.1. Modelos heurísticos basados en diccionarios	10
2.5.2. Naive Bayes y modelos estadísticos	11
2.5.3. Modelos basados en Transformers: BERT	12
2.6. Medición del desempeño	14
3. Corpus de Portuñol	17
3.1. Origen y características del corpus	17
3.2. Conversión a formato tabular	17
3.3. Preprocesamiento	18
3.3.1. Etapas del pipeline	18
3.3.2. Ejemplo antes y después	19
3.4. Anotación y etiquetas	20
3.4.1. Esquema de etiquetas	20
3.4.2. Tokenización y consideraciones	21
3.5. Estadísticas descriptivas	22
3.6. Particionado en <i>train/dev/test</i>	22
4. Experimentos	24
4.1. Heurísticas basadas en reglas	25
4.1.1. Descripción del enfoque	25
4.1.2. Resultados	27
4.2. Naive Bayes	29
4.2.1. Descripción del enfoque	29
4.2.2. Configuración y validación	29
4.2.3. Resultados	30

4.3. BERT	31
4.3.1. Descripción del enfoque	31
4.3.2. Selección del modelo BERT	32
4.3.3. Análisis de la tokenización	33
4.3.4. Configuración y validación	34
4.3.5. Resultados	34
4.4. Comparación de enfoques	38
4.4.1. Desempeño global	38
4.4.2. Comparación por categoría	39
4.4.3. Discusión general	40
4.4.4. Reflexión final	41
5. Conclusiones y trabajo futuro	42
5.1. Conclusiones	42
5.2. Trabajo futuro	43
Referencias	45

Capítulo 1

Introducción

El *portuñol*, también denominado *portugués fronterizo* o *portugués uruguayo*, es una variedad lingüística híbrida que surge del contacto prolongado entre el español y el portugués en comunidades fronterizas del norte de Uruguay, especialmente en Rivera y Artigas. Estas zonas constituyen espacios privilegiados de contacto lingüístico y cultural, donde se desafía la ideología del “un país, una lengua” y donde los hablantes construyen identidades locales a través de la alternancia de códigos (Carvalho, 2014).

A pesar de las políticas de hispanización implementadas históricamente en Uruguay, el portugués se ha mantenido como lengua de transmisión comunitaria, configurando un escenario de bilingüismo estable. Estudios sociolingüísticos han demostrado que, lejos de constituir un sistema fusionado, el portuñol mantiene continuidades con el portugués brasileño y divergencias con el español (Carvalho, 2007, 2016). Asimismo, investigaciones recientes muestran que la alternancia de lenguas cumple funciones pragmáticas específicas en la interacción social, como la construcción de humor o solidaridad (Carvalho y Christoffersen, 2022).

En lingüística, el término cambio de código (*code-switching*) se utiliza para describir el uso alternado de dos o más lenguas dentro de un mismo discurso, fenómeno que, según (A. K. Joshi, 1982), puede producirse incluso varias veces dentro de una misma oración. Este tipo de alternancia representa un desafío para los sistemas de procesamiento del lenguaje natural (PLN), dado que la mayoría de los modelos están entrenados sobre corpus que no presentan cambio de código entre lenguas (textos monolingües). En este contexto, el portuñol representa un desafío para el procesamiento automático debido a su carácter híbrido y a la variabilidad en los usos, al combinar rasgos lingüísticos del español y del portugués en distintos niveles del discurso. Dificultades similares han sido documentadas en otros escenarios de alternancia de códigos en Sudamérica, como en el contacto guaraní-español, donde la variabilidad léxica y estructural de las producciones mixtas plantea retos comparables para el PLN (Chiruzzo y cols., 2023).

El presente trabajo parte de un corpus existente de portuñol, del cual se realizó la anotación manual de un subconjunto, con el objetivo de habilitar

experimentos de clasificación automática de tokens.

El objetivo general de este proyecto es evaluar diferentes metodologías de procesamiento automático para la identificación de categorías lingüísticas en portugués. Entre los objetivos específicos se incluyen:

- Anotación de tokens en corpus con code-switching mediante etiquetas de lengua o mezcla (*es*, *pt*, *mix*, *ne*, *foreign*, *other*).
- Implementación de un modelo de referencia (*baseline*) heurístico basado en diccionarios.
- Desarrollo y evaluación de modelos probabilísticos como Naive Bayes; la exploración de arquitecturas avanzadas de tipo *transformer* (BERT).
- Análisis del desempeño de los modelos en la clasificación de tokens híbridos (*mix*), que representan el mayor desafío.

El trabajo presenta como resultado un corpus de portugués anotado manualmente, correspondiente a un subconjunto de un corpus mayor previamente existente, que permite la realización de experimentos de clasificación a nivel de token, así como una comparación sistemática de enfoques heurísticos, probabilísticos y basados en aprendizaje profundo. Los resultados evidencian las limitaciones y ventajas de cada metodología, con especial atención a la dificultad que implica la detección de tokens híbridos. Este aporte constituye una primera línea base para futuras investigaciones sobre portugués en PLN.

En términos de conclusiones, este trabajo busca aportar al campo del PLN con un recurso novedoso y con análisis aplicados a una variedad poco explorada, al tiempo que se integra en la línea de investigación sociolingüística desarrollada por [Carvalho \(2022\)](#) y otros autores. De esta forma, contribuye tanto a la comprensión del bilingüismo estable en el norte de Uruguay como a la generación de herramientas aplicables en educación, documentación lingüística y análisis cultural.

A modo de síntesis metodológica, la Figura 1.1 resume el flujo completo de trabajo seguido en esta tesis, desde la preparación y limpieza de los datos, pasando por la anotación manual y el particionado del corpus, hasta la etapa experimental y la evaluación de los modelos. La figura permite visualizar de manera integrada las principales decisiones y etapas del proceso.

El documento se organiza de la siguiente manera: en el Capítulo 2 se presenta la revisión de antecedentes sobre el contacto entre español y portugués y estudios previos en PLN sobre variedades mixtas; en el Capítulo 3 se describe la metodología utilizada, con énfasis en el proceso de anotación del corpus; en el Capítulo 4 se detallan los experimentos realizados y los modelos implementados, se analizan y discuten los resultados obtenidos; y en el Capítulo 5 se presentan las conclusiones generales junto con posibles líneas de investigación futura.

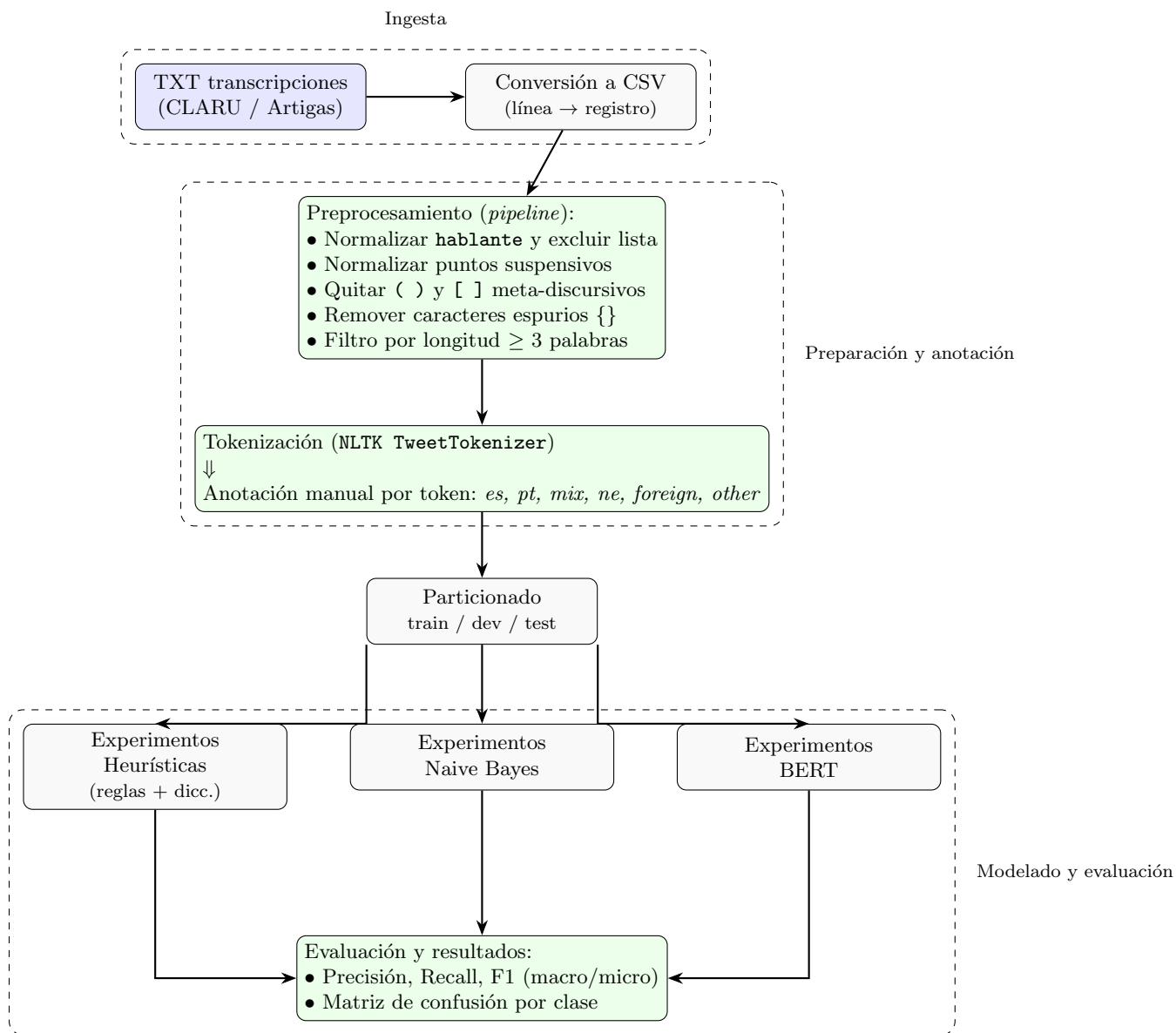


Figura 1.1: Diagrama de flujo del pipeline: desde TXT (CLARU) hasta los experimentos (Heurísticas, Naive Bayes, BERT) y evaluación.

Capítulo 2

Marco Teórico

En este capítulo se introducen los conceptos necesarios para comprender el presente trabajo, así como los enfoques y herramientas empleados en el análisis. Se abordan temas relacionados con el portugués uruguayo, la alternancia de códigos, la detección automática de idioma en PLN y los principales métodos de clasificación utilizados, desde heurísticas y modelos estadísticos hasta arquitecturas profundas basadas en *transformers*.

2.1. El portugués uruguayo y el portuñol

El portugués uruguayo, también denominado portugués fronterizo, es la variedad hablada en el norte de Uruguay, especialmente en Rivera y Artigas, como resultado del contacto prolongado con el español. *Estudios variacionistas*¹ han mostrado que se trata de una variedad del portugués brasileño que mantiene continuidades estructurales con esta lengua, pero que está fuertemente influida por el contacto con el español (Carvalho, 2007, 2016).

En este contexto, el término *portuñol* se utiliza para describir el repertorio híbrido que los hablantes emplean en interacciones fronterizas. Más allá de las percepciones de “mezcla”, las investigaciones de Carvalho (2014, 2022) demuestran que el portuñol conserva diferencias claras respecto al español y que la alternancia de códigos responde a patrones sistemáticos. Además, la alternancia cumple funciones pragmáticas, como la producción de humor, ironía o solidaridad, según muestran Carvalho y Christoffersen (2022).

La Figura 2.1 muestra una propuesta de distribución de las principales variedades dialectales del español en Uruguay, incluyendo una zona identificada como *fronterizo portuñol* en el norte del país. Esta región comprende, entre otros, los departamentos de Rivera y Artigas, y coincide con el área de estudio abordada en esta tesis.

¹Una corriente de la sociolingüística que analiza cómo las variables sociales (como edad, género o clase social) se relacionan con la variación lingüística.

El mapa permite contextualizar geográficamente el portugués uruguayo dentro del continuo dialectal del español en Uruguay, mostrando que se trata de una variedad localizada y diferenciada, asociada a la frontera con Brasil y al contacto prolongado con el portugués. Esta delimitación resulta relevante para comprender las particularidades lingüísticas del corpus analizado y los desafíos que plantea su modelización computacional.



Figura 2.1: Distribución de variedades dialectales del español en Uruguay, incluyendo la región identificada como *fronterizo portugués* en el norte del país. Fuente: Adaptado de https://es.wikipedia.org/wiki/Lenguas_de_Uruguay.

2.2. Alternancia de códigos (Code-switching)

La alternancia de códigos (code-switching) consiste en el uso de dos lenguas dentro de un mismo discurso o conversación. Aunque a menudo se percibe como interferencia, numerosas investigaciones han demostrado que se trata de un fenómeno estructurado y funcional, condicionado por factores sociales, discursivos y pragmáticos (Poplack, 1980; Myers-Scotton, 1993; Bullock y Toribio, 2009).

Este fenómeno ha sido ampliamente documentado en distintas comunidades bilingües del mundo —por ejemplo, entre inglés y español en Estados Unidos, francés y árabe en el Magreb, o hindi e inglés en India—, lo que evidencia su carácter generalizado y su relevancia para el estudio del contacto de lenguas (Muysken, 2000; Gardner-Chloros, 2009). Aunque la mayoría de los trabajos iniciales se centraron en su descripción sociolingüística, también se han propuesto enfoques formales y computacionales que buscan modelar la alternancia dentro de una misma oración (por ejemplo, A. K. Joshi (1982)).

En el caso del español y portugués, Carvalho (2022) sistematiza los escenarios de contacto en Sudamérica y resalta que el norte de Uruguay constituye un ejemplo único de bilingüismo estable en el siglo XXI. Fenómenos de contacto entre el español y el portugués también se observan en otras regiones del mundo, como en Galicia (España), donde el gallego —una lengua estrechamente emparentada con el portugués— coexiste con el castellano, dando lugar a formas híbridas conocidas como *castrapo*². Estos casos muestran que los procesos de mezcla y alternancia entre lenguas romances pueden emerger en distintos contextos geográficos, aunque con características sociolingüísticas particulares.

En los últimos años, el estudio computacional del code-switching ha cobrado impulso en diversas competiciones lingüísticas, como el guaraní-español en la competencia GUA-SPA @ IberLEF 2023 (Chiruzzo y cols., 2023), donde se abordó la identificación automática de lengua y el reconocimiento de entidades en textos mixtos. Estos avances reflejan un interés creciente por el tratamiento computacional del cambio de código en variedades sudamericanas.

Además de las transiciones entre lenguas dentro de una oración, el code-switching también involucra la aparición de **entidades nombradas** (*named entities*, NE), como nombres propios de personas, lugares, instituciones o marcas. Estas unidades presentan un comportamiento particular en contextos bilingües: conservan su forma léxica estable, pero pueden adaptarse fonológica u ortográficamente a la lengua dominante en el discurso (*p. ej.*, *Brasil* / *Brazil*). En el procesamiento del lenguaje natural (PLN), las NE se tratan como una categoría independiente, ya que su identificación requiere distinguir entre vocabulario común y unidades semánticamente fijas (Jurafsky y Martin, 2023). Su

² *Castrapo* es el término que designa las formas híbridas resultantes del contacto entre el gallego y el castellano en el noroeste de España. Se trata de un fenómeno sociolingüístico asociado a situaciones de bilingüismo y contacto prolongado entre dos lenguas romances estrechamente emparentadas, comparable al portuñol por la presencia de mezclas estructurales y léxicas. Véase: Centro Virtual Cervantes, <https://cvc.cervantes.es/obref/congresos/valladolid/ponencias/unidad.diversidad.del.espanol/4.el.espanol.en.contacto/porto.a.htm>.

tratamiento diferenciado resulta esencial en corpus multilingües o de alternancia de códigos, donde pueden funcionar como puntos de transición entre lenguas o ser confundidas con tokens extranjeros. Por esta razón, muchos esquemas de anotación —incluido el adoptado en esta tesis— asignan una etiqueta específica (*ne*) para representar estas entidades y analizar su distribución en textos bilingües.

2.2.1. Tokens híbridos (*mix*) y su complejidad para el PLN

En contextos de alternancia de códigos, además de los cambios entre segmentos completos de lengua, pueden aparecer **tokens híbridos** o *mix*, que combinan en una misma forma rasgos léxicos, morfológicos u ortográficos de las lenguas en contacto.

Desde el punto de vista lingüístico, los tokens híbridos se sitúan en un *continuum* entre la alternancia interoracional —donde cada segmento pertenece a una lengua claramente identificable— y la fusión morfológica, donde la frontera entre sistemas se difumina. Este fenómeno ha sido documentado en distintos contextos de contacto lingüístico (Muysken, 2000; Gardner-Chloros, 2009), y representa uno de los principales retos para el procesamiento automático.

Para el PLN, los tokens *mix* resultan especialmente problemáticos por varias razones. En primer lugar, suelen estar ausentes de los diccionarios monolingües y, por tanto, son tratados como palabras desconocidas. Además, los tokenizadores basados en subpalabras tienden a dividirlos en múltiples unidades, lo que genera representaciones fragmentadas y de difícil interpretación. Esta sobresegmentación —reflejada en métricas como la *subword fertility* y la *subword proportion* (Rust y cols., 2021)— limita la capacidad de los modelos para capturar su significado contextual. Finalmente, su ambigüedad lingüística los vuelve difíciles de clasificar incluso para modelos neuronales avanzados, que requieren información contextual más amplia para decidir la etiqueta correcta.

En síntesis, los tokens híbridos condensan los desafíos del *code-switching*: combinan variación morfológica, interferencia ortográfica y escasez de datos, factores que explican su bajo rendimiento relativo en tareas de detección automática de idioma (Solorio y cols., 2014; Aguilar, Kar, y Solorio, 2020).

2.3. Lenguas de escasos recursos en PLN

En el procesamiento del lenguaje natural, se denomina **lenguas de escasos recursos** (*low-resource languages*, LRLs) a aquellas para las cuales existen pocos o nulos recursos digitales, tales como corpus anotados, herramientas lingüísticas, modelos preentrenados o materiales de referencia en línea (P. Joshi, Santy, Budhiraja, Bali, y Choudhury, 2020). Esta carencia limita la posibilidad de desarrollar sistemas robustos, comparables entre lenguas y reproducibles, lo que genera una brecha persistente entre los idiomas mayoritarios y los minoritarios.

En los últimos años, la investigación en PLN ha propuesto diversas estrategias para mitigar esta escasez, entre ellas el *transfer learning* multilingüe, el uso de modelos de propósito general (como mBERT) y la *data augmentation* o generación sintética de ejemplos. Dentro de esta última línea, [Lucas y cols. \(2024\)](#) presentan un enfoque basado en reglas gramaticales para aumentar datos en un sistema de traducción automática Guaraní-Español. Su método demuestra que la generación controlada de ejemplos puede mejorar significativamente el desempeño de los modelos en lenguas con muy poca disponibilidad de datos, ofreciendo un camino replicable para otras variedades de contacto.

El caso del **portuñol** se inscribe en este mismo marco de problemáticas. Al no contar con corpus públicos ni herramientas lingüísticas, constituye una variedad de *muy bajos recursos*, además de presentar particularidades estructurales derivadas del contacto entre el español y el portugués. Estas características dificultan la aplicación directa de modelos entrenados sobre lenguas monolingües y obligan a desarrollar estrategias específicas de adaptación o transferencia.

En este sentido, estudiar el portuñol no solo permite avanzar en la descripción lingüística de una variedad fronteriza, sino también aportar evidencia sobre cómo abordar el PLN en contextos de escasez de datos, un desafío compartido por numerosas lenguas minoritarias y en contacto a nivel global.

2.4. Corpus existentes para el portuñol

A diferencia de otras lenguas en contacto, el portuñol carece de corpus lingüísticos públicos o conjuntos de datos disponibles para investigación en procesamiento del lenguaje natural (PLN). Hasta la fecha, no existen recursos estandarizados ni bases de datos con anotaciones lingüísticas —como etiquetas de idioma o segmentaciones a nivel de token— que permitan entrenar o evaluar modelos automáticos. Esta ausencia de recursos refleja la doble condición del portuñol: por un lado, su estatus como variedad oral, local y no institucionalizada; por otro, su pertenencia al conjunto de las lenguas y variedades de *muy bajos recursos*, donde la producción de datos depende casi exclusivamente de esfuerzos individuales o académicos aislados.

El único proyecto de documentación sistemática conocido es el **Corpus Lingüístico de Artigas RurUrbano (CLARU)**, dirigido por [Garrido Sánchez-Andrade \(2025\)](#). Este corpus se centra exclusivamente en la ciudad de Artigas y su entorno rural, con el propósito de registrar la variedad fronteriza hablada en la región norte de Uruguay. CLARU está compuesto por treinta entrevistas y quince conversaciones espontáneas entre los mismos informantes, grabadas durante los meses de noviembre y diciembre de 2021. Las transcripciones, realizadas entre 2022 y 2024 por la investigadora responsable, ofrecen un material valioso para la descripción sociolingüística de la variedad, aunque todavía no está disponible públicamente para análisis computacional. Se prevé su publicación en 2025 a través de la plataforma *SWISSUbase*³.

³<https://www.swissubase.ch/en/>

Fuera del ámbito académico, parte del portuñol uruguayo se conserva y difunde en la literatura contemporánea, especialmente en la obra del escritor **Fabián Severo**, cuyas publicaciones —como *Noite nu Norte*, *Viento de nadie* o *Virala-ta*— exploran el habla fronteriza desde una perspectiva artística y reivindicativa. Aunque estos textos no constituyen un corpus lingüístico en sentido técnico, sí representan una fuente escrita significativa para el estudio del portuñol como fenómeno de contacto, y han sido empleados como material de referencia en investigaciones sobre variación y bilingüismo.

En síntesis, el estado actual de los recursos disponibles evidencia una **escasez de corpus lingüísticos y materiales digitalizados** sobre el portuñol. La mayoría de los datos existentes son de naturaleza cualitativa y no han sido adaptados para su uso en tareas de PLN. Este vacío motiva la construcción y anotación de nuevos conjuntos de datos, como los desarrollados en el marco de esta tesis, orientados a avanzar en la caracterización automática de la variedad.

2.5. Detección automática de idioma en PLN

La *detección automática de idioma* (*Language Identification*, LID) es una tarea fundamental en el procesamiento del lenguaje natural (PLN). En corpus monolingües extensos, los sistemas alcanzan altos niveles de precisión —frecuentemente superiores al 99 % en condiciones controladas— gracias al uso de modelos estadísticos y neuronales entrenados sobre grandes volúmenes de datos (Jauhiainen, Lui, Zampieri, Baldwin, y Lindén, 2019; Lui y Baldwin, 2012; Brown y cols., 2020). Sin embargo, en escenarios de lenguas en contacto y alternancia de códigos, la tarea presenta dificultades adicionales, especialmente en tokens híbridos o ambiguos, donde los límites entre las lenguas no son nítidos (Solorio y cols., 2014; Aguilar y cols., 2020; Chiruzzo y cols., 2023).

En una revisión extensa, Jauhiainen y cols. (2019) clasifican los métodos de LID en tres grandes grupos: (i) enfoques heurísticos basados en reglas y diccionarios, (ii) modelos estadísticos como n-gramas y Naive Bayes, y (iii) modelos neuronales modernos. De manera similar, Hughes, Baldwin, Bird, Nicholson, y MacKinlay (2006) y Solorio y cols. (2014) destacan la evolución desde aproximaciones iniciales basadas en listas de palabras hacia enfoques probabilísticos y, más recientemente, a arquitecturas de aprendizaje profundo.

En contextos multilingües, la identificación de idioma requiere manejar no solo límites entre palabras, sino también fenómenos de *code-switching* y *code-mixing*, en los que pueden coexistir varias lenguas dentro de una misma oración o incluso en una sola palabra. Estudios recientes en competencias internacionales, como LinCE (Aguilar y cols., 2020) o GUA-SPA (Chiruzzo y cols., 2023), muestran que la precisión de los sistemas disminuye significativamente en comparación con escenarios monolingües, lo que evidencia la complejidad de la tarea en datos de contacto lingüístico.

Dada la relevancia de estos enfoques para el trabajo presentado en este documento, se describirá cada uno de ellos a continuación.

2.5.1. Modelos heurísticos basados en diccionarios

Los enfoques heurísticos constituyen uno de los métodos más antiguos en la detección automática de idioma. Se basan en reglas explícitas y en el uso de diccionarios o listas de palabras por lengua, que permiten asignar una etiqueta a cada token en función de coincidencias directas. Por su simplicidad e interpretabilidad, estos modelos han sido ampliamente utilizados como línea de base en estudios iniciales (Jauhiainen y cols., 2019).

La lógica general consiste en verificar si una palabra aparece en un listado asociado a una lengua determinada y, en caso afirmativo, asignar dicha etiqueta. Cuando las palabras no aparecen en ningún diccionario, suelen clasificarse como desconocidas, entidades nombradas o extranjerismos, según el contexto.

Estos métodos tienen como ventaja la transparencia y la facilidad de implementación, ya que no requieren entrenamiento ni grandes cantidades de datos. Sin embargo, su cobertura está limitada al vocabulario incluido en los diccionarios, lo que dificulta la clasificación de variantes dialectales y errores ortográficos.

En este trabajo, estos métodos se aplican a la variedad lingüística objeto de estudio: el portugués. En este contexto, los enfoques heurísticos enfrentan dificultades adicionales para identificar correctamente tokens híbridos (*mix*), característicos de esta alternancia de códigos, lo que reduce su efectividad en escenarios complejos de contacto lingüístico (Hughes y cols., 2006).

2.5.2. Naive Bayes y modelos estadísticos

El clasificador *Naive Bayes* es un modelo probabilístico basado en el teorema de Bayes, que estima la probabilidad de que un elemento pertenezca a una clase determinada en función de sus características observadas (Mitchell, 1997). Bajo el supuesto de independencia condicional entre las características, la probabilidad posterior se define como:

$$P(C_k | \mathbf{x}) = \frac{P(C_k) P(\mathbf{x} | C_k)}{P(\mathbf{x})}, \quad (2.1)$$

donde C_k representa una clase (por ejemplo, el idioma), y $\mathbf{x} = (x_1, x_2, \dots, x_n)$ son las características del token o texto analizado.

El supuesto *naive* consiste en considerar que los atributos son condicionalmente independientes dado C_k , lo que permite factorizar:

$$P(\mathbf{x} | C_k) = \prod_{i=1}^n P(x_i | C_k). \quad (2.2)$$

Así, la predicción se obtiene mediante la regla de máxima verosimilitud a posteriori:

$$\hat{C} = \arg \max_{C_k} P(C_k) \prod_{i=1}^n P(x_i | C_k). \quad (2.3)$$

Pese a la simplicidad del supuesto, Naive Bayes ha mostrado un desempeño competitivo en tareas de clasificación de texto y se utiliza frecuentemente como modelo base en PLN. En la detección de idioma, el modelo se entrena a partir de *representaciones estadísticas del texto*, como la distribución de caracteres o las secuencias de unidades conocidas como *n-gramas*.

Un *n-grama* es una secuencia continua de n unidades extraídas de un texto, que pueden ser letras, sílabas o palabras, siendo los *n-gramas* de caracteres los más utilizados para la detección automática de idiomas. Por ejemplo, para $n = 3$, el término *casa* genera los trigramas *cas* y *asa*. Los modelos basados en *n-gramas* capturan regularidades estadísticas propias de cada lengua y permiten estimar distribuciones de probabilidad sobre secuencias (Jurafsky y Martin, 2023). Esto los hace especialmente útiles para la identificación de idioma y otros problemas de modelado lingüístico.

Una dificultad común en la aplicación de modelos estadísticos es la presencia de eventos no observados —por ejemplo, palabras o combinaciones que no aparecen en el conjunto de entrenamiento—, lo que puede llevar a probabilidades nulas. Para mitigar este problema se utilizan técnicas de **suavización**, entre

ellas el *m-estimador*, que ajusta las probabilidades condicionales otorgando un pequeño peso a los eventos no vistos. Este mecanismo mejora la capacidad de generalización del modelo, especialmente frente a clases minoritarias o tokens infrecuentes, como los etiquetados como *mix* en el corpus de portugués.

Formulación de la tarea. En el contexto de esta tesis, el modelo *Naive Bayes* se aplica a la clasificación de tokens individuales según su idioma. Cada token se representa a través de una distribución de frecuencias de n-gramas de caracteres, lo que permite capturar información fonológica y ortográfica sin requerir segmentación lingüística compleja. El modelo estima la probabilidad de cada etiqueta posible —*es*, *pt*, *mix*, *foreign*, *ne* u *other*— y asigna al token aquella que maximiza la probabilidad posterior. Esta formulación probabilística, sencilla y transparente, sirve como línea de base interpretativa para comparar con los modelos neuronales descritos en las siguientes secciones.

2.5.3. Modelos basados en Transformers: BERT

La arquitectura *Transformer* introdujo el mecanismo de *self-attention*, revolucionando el PLN al permitir modelar dependencias a largo plazo de manera eficiente (Vaswani y cols., 2017). Sobre esta base, Devlin, Chang, Lee, y Toutanova (2019) desarrollaron *BERT* (Bidirectional Encoder Representations from Transformers), entrenado mediante enmascarado de palabras y predicción de la siguiente oración. Este modelo genera representaciones profundas y bidireccionales de los tokens, lo que le otorga un alto rendimiento en tareas de clasificación y comprensión del lenguaje.

En escenarios multilingües, variantes como *bert-base-multilingual-cased* permiten representar múltiples lenguas en un mismo espacio vectorial, lo que las hace especialmente útiles para la identificación de tokens en contextos de alternancia de códigos. En este trabajo se explora el uso de **BERT multilingüe** y de variantes específicas en español y portugués, con el objetivo de evaluar su desempeño en la clasificación automática de tokens del corpus de portugués.

Diversos estudios han demostrado que los modelos basados en BERT alcanzan resultados sobresalientes en tareas de clasificación, superando ampliamente a los enfoques estadísticos tradicionales. En particular, Deka y cols. (2022) aplicaron un modelo BERT para la detección de idioma en textos con *code-switching* y observaron mejoras significativas respecto de los métodos basados en n-gramas. Estos resultados respaldan la efectividad de BERT para representar información contextual y resolver tareas lingüísticas complejas en entornos multilingües.

Entrenamiento, *fine-tuning* y *early stopping*. El proceso de **entrenamiento** de modelos neuronales consiste en ajustar los parámetros internos de la red para minimizar una función de pérdida sobre un conjunto de datos. En el caso de modelos preentrenados como BERT, no se parte de una red aleatoria, sino de pesos previamente optimizados sobre grandes corpus multilingües. El ***fine-tuning*** (ajuste fino) consiste en continuar ese entrenamiento sobre un conjunto

de datos específico para una tarea determinada —en este caso, la clasificación de tokens por idioma—, adaptando las representaciones generales del modelo al dominio y etiquetas del nuevo problema.

Durante el entrenamiento se aplica el criterio de *early stopping*, que detiene el proceso cuando la métrica de validación (por ejemplo, la F1 macro) deja de mejorar tras un número determinado de épocas consecutivas. Esta técnica previene el sobreajuste (*overfitting*) y favorece la generalización del modelo, evitando continuar el ajuste cuando la red comienza a memorizar el conjunto de entrenamiento.

Formulación de la tarea. En esta tesis, el modelo BERT se utiliza bajo el paradigma de **clasificación por token** mediante la clase `BertForTokenClassification` de la biblioteca *Transformers* de Hugging Face.⁴, que proyecta las representaciones contextuales de cada token hacia el espacio de etiquetas lingüísticas. Las instancias del corpus se reconstruyen agrupando los tokens que pertenecen a un mismo texto, de modo que cada secuencia de entrada corresponde a un fragmento coherente identificado por `id_texto`. Posteriormente, las secuencias se tokenizan y se preparan para el modelo mediante un esquema de alineamiento *subtoken-aware*: solo el primer subtoken de cada palabra recibe la etiqueta correspondiente, mientras que los subtokens restantes se enmascaran para que no contribuyan al cálculo de la pérdida. El modelo se entrena bajo un enfoque de *fine-tuning* supervisado y se evalúa a nivel de token utilizando métricas de **precisión, exhaustividad y F1 macro**, así como matrices de confusión que permiten analizar los errores por clase.

A diferencia de la clasificación a nivel de secuencia —donde se utiliza el token especial [CLS]—, en esta tarea se emplea directamente la representación contextual de cada token individual para predecir su etiqueta. Esta formulación permite modelar la alternancia de lenguas dentro de una misma oración y capturar información contextual relevante para los tokens híbridos o ambiguos.

⁴https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForTokenClassification

2.6. Medición del desempeño

La evaluación de modelos de clasificación requiere métricas cuantitativas que permitan comparar diferentes enfoques. Las más utilizadas en aprendizaje supervisado son la *precisión* (*precision*), la *exhaustividad* (*recall*) y la *F1-score*.

- **Precisión (Precision)**: mide la proporción de instancias correctamente clasificadas como positivas respecto del total de predicciones positivas realizadas por el modelo:

$$\text{Precision} = \frac{TP}{TP + FP}$$

donde TP representa los verdaderos positivos y FP los falsos positivos.

- **Exhaustividad (Recall)**: indica la proporción de instancias positivas que el modelo logra identificar correctamente sobre el total de instancias reales de esa clase:

$$\text{Recall} = \frac{TP}{TP + FN}$$

donde FN corresponde a los falsos negativos.

- **F1-score**: corresponde a la media armónica entre la precisión y la exhaustividad, y proporciona un balance entre ambas métricas:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Cuando se trabaja con múltiples clases, las métricas pueden calcularse de forma *macro*, *micro* o *ponderada* (*weighted*), dependiendo de si se promedian las métricas de cada clase por igual, según la proporción de instancias, o globalmente sobre todos los ejemplos.

Además, la **matriz de confusión** es una herramienta útil para visualizar los aciertos y errores de un modelo, mostrando las clases predichas frente a las reales. Cada celda M_{ij} contiene el número de ejemplos cuya clase real es i y que fueron predichos como j . Esta representación permite identificar con claridad qué clases son confundidas entre sí y dónde se concentran los errores.

En contextos de *code-switching*, estas métricas resultan especialmente relevantes, ya que permiten analizar qué categorías lingüísticas son más difíciles de predecir. Estudios previos han mostrado que la clase *mix* suele presentar menor desempeño en comparación con las etiquetas puramente monolingües, debido a su ambigüedad y variabilidad (Chiruzzo y cols., 2023).

Si bien la *accuracy* ofrece una medida global del número total de aciertos, su valor puede resultar engañoso cuando las clases están desbalanceadas, ya que las categorías mayoritarias dominan el promedio general. En estos casos, métricas como la **F1-score macro** resultan más expresivas, al otorgar el mismo peso

a cada clase sin importar su frecuencia. De esta manera, se obtiene una evaluación más equitativa del rendimiento del modelo sobre todas las categorías, incluyendo las minoritarias. Asimismo, aunque el *recall* es útil para identificar qué proporción de ejemplos positivos logra recuperar el modelo, su interpretación aislada puede ser incompleta: una alta exhaustividad acompañada de baja precisión implica un gran número de falsos positivos. La F1-score, al combinar ambas dimensiones, permite equilibrar estos efectos y proporcionar una estimación más robusta del desempeño real en escenarios con distribución desigual de clases.

Métricas de evaluación de tokenizadores

En los modelos basados en *Transformers*, como BERT, la tokenización no se realiza a nivel de palabras completas, sino mediante técnicas de segmentación subléxica que dividen las palabras en unidades más pequeñas denominadas *subtokens*. Estos tokenizadores operan a nivel subléxico y se entrenan automáticamente a partir de grandes corpus textuales, utilizando métodos estadísticos que buscan aprender un inventario de subtokens capaz de equilibrar cobertura léxica y eficiencia representacional.

Cuando una lengua o variedad lingüística está escasamente representada en los datos utilizados para entrenar el tokenizador, es frecuente que muchas palabras no sean reconocidas como unidades completas y deban fragmentarse en múltiples subtokens. Esta fragmentación excesiva puede degradar la representación semántica de las palabras y dificultar el aprendizaje contextual del modelo.

En este sentido, la calidad del tokenizador influye directamente en el desempeño de los modelos basados en *Transformers*, especialmente en escenarios de contacto lingüístico como el portuñol, donde abundan formas no estándar e híbridas.

Para analizar esta etapa, [Rust y cols. \(2021\)](#) proponen dos métricas que permiten cuantificar la cobertura y granularidad de la segmentación subléxica:

- **Subword fertility**: mide la cantidad promedio de sub-tokens generados por cada palabra tokenizada. Se calcula como:

$$\text{Fertility} = \frac{1}{N} \sum_{i=1}^N s_i$$

donde N es el número total de palabras en el corpus y s_i la cantidad de sub-tokens en que se divide la palabra i . Valores bajos indican una segmentación más compacta y una mejor cobertura del vocabulario.

- **Proporción de palabras continuadas (Subword proportion)**: representa la proporción de palabras que se dividen en más de un sub-token:

$$\text{Proportion} = \frac{1}{N} \sum_{i=1}^N 1(s_i > 1)$$

donde $1(\cdot)$ es una función indicadora que vale 1 si la condición se cumple y 0 en caso contrario. Una proporción baja implica que la mayoría de

las palabras se representan con un único token, lo que sugiere una mejor adecuación del tokenizador a la lengua.

Por su definición, estas métricas presentan rangos bien delimitados. La *subword fertility* tiene como valor mínimo teórico 1, correspondiente al caso en que todas las palabras se representan mediante un único subtoken, y no posee un límite superior fijo, ya que depende del grado de fragmentación subléxica. En cambio, la *subword proportion* está acotada entre 0 y 1: toma el valor 0 cuando ninguna palabra se divide en múltiples subtokens, y el valor 1 cuando todas las palabras requieren más de un subtoken.

Estas métricas permiten comparar tokenizadores en términos de su eficiencia para representar palabras completas. Valores reducidos de *fertility* y *proportion* reflejan una cobertura más adecuada del vocabulario y menor fragmentación subléxica, aspectos deseables en modelos multilingües y en contextos de *code-switching* (Rust y cols., 2021).

Capítulo 3

Corpus de Portuguol

Este capítulo describe el corpus utilizado, su procedencia, el formato original y el proceso de preparación previo a los experimentos. A diferencia de trabajos que construyen el recurso desde cero, en esta tesis se parte de un corpus existente y se realiza la anotación manual de un subconjunto del mismo, con el objetivo de habilitar su uso en tareas de clasificación automática a nivel de *token*.

3.1. Origen y características del corpus

El corpus empleado proviene del *Corpus Lingüístico de Artigas RurUrbano* (CLARU), desarrollado en la Universidad de Zúrich por la investigadora Garrido Sánchez-Andrade (2025). El corpus fue distribuido en formato `.txt` e incluye transcripciones realizadas entre noviembre y diciembre de 2021 en el departamento de Artigas, Uruguay. El material reúne dos tipos de registros:

- **Entrevistas dirigidas:** un investigador formula preguntas a hablantes de portuguol sobre aspectos de su vida cotidiana y prácticas culturales.
- **Conversaciones espontáneas:** diálogos entre dos hablantes de portuguol, sin intervención del entrevistador, que capturan fenómenos de alternancia de códigos en un contexto natural.

El CLARU se compone de treinta entrevistas y quince conversaciones, registradas tanto en el medio urbano (ciudad de Artigas) como en el rural (La Guayubira y La Estiba). En total, el conjunto considerado para esta tesis comprende **46 archivos** y **16 935 líneas de texto transcrito**.

3.2. Conversión a formato tabular

Para facilitar el procesamiento, las transcripciones originales en formato `.txt` fueron convertidas a un archivo `CSV`, donde cada línea representa un registro independiente. El nuevo formato incluye al menos dos columnas principales:

- **hablante**: identificador del hablante correspondiente a la línea.
- **texto**: contenido textual transcrito.

Durante la conversión se preservó el orden de aparición de las líneas y la segmentación original de las transcripciones. Esto permite mantener la correspondencia entre las secciones textuales y su contexto de aparición dentro de cada archivo fuente. Por ejemplo, si una entrevista incluía intervenciones alternadas entre hablante y entrevistador, el orden relativo se conserva, lo que facilita el posterior análisis de alternancia de códigos en el discurso.

A modo ilustrativo, la Tabla 3.3 muestra un fragmento simplificado del formato tabular resultante, con algunas líneas representativas del corpus.

hablante	texto
E	eh... NP tu tem quantos anos?
H1	quarenta e quatro
E	quarenta e quatro. eh é.. de Artigas...
H1	de Artigas

Tabla 3.1: Ejemplo ilustrativo del formato tabular utilizado para representar las transcripciones.

3.3. Preprocesamiento

Antes de la anotación manual, se aplicó un *pipeline* determinista de limpieza y normalización para obtener un conjunto textual consistente y sin ruido. El objetivo fue estandarizar la estructura de los archivos y eliminar marcas no lingüísticas que pudieran interferir en las etapas de tokenización y anotación.

3.3.1. Etapas del pipeline

1. **Normalización de hablante y exclusión por lista**: se eliminaron los registros correspondientes al entrevistador y a hablantes identificados en una lista de exclusión, principalmente aquellos con intervenciones sin texto. Esta depuración garantiza que las líneas procesadas correspondan efectivamente a discurso oral en portugués y no a elementos de la interacción metodológica.
2. **Normalización de puntos suspensivos**: las secuencias de dos o más puntos ($\dots$, $\dots$, $\dots$) fueron reducidas a un único punto. Esta decisión se tomó porque los puntos suspensivos, frecuentes en las transcripciones, suelen indicar pausas o vacilaciones del habla y no delimitan límites oracionales claros. Su normalización contribuye a una segmentación más estable y evita que el tokenizador los interprete como secuencias de símbolos independientes.

3. **Eliminación de contenido entre paréntesis o corchetes:** en las transcripciones se incluyen marcas meta-discursivas o paralingüísticas, tales como *(risa)*, *[ruido]*, *(pausa)*, utilizadas por la investigadora para documentar fenómenos del registro oral. Aunque estos elementos son valiosos para el análisis sociolingüístico, no forman parte del material lingüístico propiamente dicho y, por lo tanto, se eliminaron para no interferir en la segmentación y el conteo de tokens. De esta forma se preserva únicamente la secuencia verbal del hablante.
4. **Eliminación de caracteres espurios:** se identificaron símbolos aislados o artefactos de transcripción, como {, }, probablemente introducidos durante el proceso de codificación de texto. Estos caracteres fueron eliminados ya que no aportan información lingüística y pueden alterar la tokenización posterior.
5. **Filtro por longitud mínima:** se descartaron líneas con menos de tres palabras. La decisión de establecer este umbral se basó en un análisis exploratorio realizado con la librería `TweetTokenizer` de NLTK, que permitió observar que las líneas muy cortas correspondían casi siempre a interjecciones aisladas (“*ah*”, “*si*”, “*claro*”) o fragmentos incompletos de diálogo. Estas secuencias, además de carecer de contexto, generan problemas al calcular distribuciones de frecuencia o al aplicar modelos supervisados, ya que no aportan información suficiente para inferir rasgos de alternancia de códigos. El umbral de tres palabras fue, por tanto, un compromiso entre mantener la representatividad del corpus y garantizar la informatividad mínima de cada segmento.

En conjunto, este *pipeline* permitió depurar y estandarizar los textos, reduciendo la ambigüedad y mejorando la calidad del material que posteriormente se utilizó para la tokenización y anotación a nivel de *token*. El diseño de estas etapas buscó equilibrar la limpieza formal con la preservación del contenido lingüístico relevante: las exclusiones por hablante y las líneas de longitud mínima redujeron el ruido y evitaron sesgos en la anotación, mientras que la normalización de puntuación y la eliminación de marcas paralingüísticas estabilizaron la segmentación en *tokens*. Estas decisiones contribuyeron a obtener un corpus coherente, homogéneo y adecuado para el entrenamiento de modelos de detección de idioma, minimizando la influencia de variaciones ortográficas o símbolos ajenos al fenómeno lingüístico (Jauhainen y cols., 2019; Hughes y cols., 2006).

3.3.2. Ejemplo antes y después

La Tabla 3.2 ilustra el efecto del *pipeline* sobre una línea típica del corpus.

Antes (texto original)	Después (texto limpio)
ah... uma raiada e tá (risa)}	ah. uma raiada e tá

Tabla 3.2: Ejemplo de limpieza: supresión de marcas meta-discursivas, normalización de puntos suspensivos y remoción de caracteres espurios.

3.4. Anotación y etiquetas

Sobre el conjunto limpio se efectuó una **anotación manual parcial** a nivel de *token*. El objetivo fue disponer de un subconjunto representativo del corpus original para el entrenamiento y evaluación de modelos de detección automática de idioma. En total se anotaron **31 054 tokens**, correspondientes a aproximadamente el **13.37 %** del total del corpus (**232 195 tokens**). Esta porción resultó suficiente para cubrir la variación entre hablantes, géneros discursivos y fenómenos de alternancia de códigos.

3.4.1. Esquema de etiquetas

La anotación se realizó empleando seis categorías lingüísticas, inspiradas en las convenciones utilizadas en tareas internacionales de *code-switching* como CALCS y GUA-SPA (Solorio y cols., 2014; Chiruzzo y cols., 2023). Las etiquetas y su descripción son las siguientes:

- **es**: token en español estándar.
- **pt**: token en portugués estándar.
- **mix**: token híbrido o morfológicamente mixto que combina rasgos de ambas lenguas.
- **ne**: *named entity*, entidades con nombre propio (personas, lugares, instituciones). En PLN se reconocen como una categoría independiente por su comportamiento léxico estable y su ambigüedad interlingüística.
- **foreign**: tokens provenientes de lenguas distintas del español y portugués (por ejemplo, *okay*, *bye*).
- **other**: signos de puntuación y otros símbolos no lingüísticos.

La distinción entre *es*, *pt* y *mix* permite capturar distintos grados de alternancia de códigos, mientras que *ne* y *foreign* aíslan elementos que suelen generar confusión en los sistemas automáticos de detección de idioma. El fenómeno de *code-switching* se manifiesta tanto entre palabras consecutivas de diferentes lenguas como dentro de una misma palabra (caso de los tokens *mix*), lo que justifica la necesidad de una anotación a nivel de *token*.

El proceso de anotación se efectuó en una planilla CSV estructurada con las columnas: `archivo`, `linea`, `hablante`, `texto_completo`, `token`, `etiqueta`, `comentario`. Cada fila representa un token individual con sus metadatos de

contexto. Esta estructura facilita la trazabilidad entre las anotaciones y el texto fuente, permitiendo futuras revisiones o ampliaciones del corpus.

3.4.2. Tokenización y consideraciones

La segmentación en tokens se realizó con la librería `TweetTokenizer` de NLTK, una herramienta ampliamente utilizada en textos informales. Su elección se basó en recomendaciones de los tutores del proyecto, quienes la habían empleado en trabajos previos con buenos resultados. El tokenizador funcionó correctamente sobre las transcripciones del portuñol, separando de forma adecuada los tokens sin generar errores relevantes.

Este tokenizador presenta ventajas para datos orales y no estandarizados, ya que maneja de manera robusta contracciones, puntuación ligera y expresiones con emoticones o risas transcritas. En el corpus del portuñol, caracterizado por variación ortográfica y fenómenos de habla espontánea, esta herramienta permitió mantener la integridad de los tokens híbridos y reducir errores de segmentación.

En síntesis, el esquema de anotación propuesto combina criterios lingüísticos y funcionales para capturar la complejidad del portuñol, equilibrando la precisión lingüística con la viabilidad computacional. El uso de un tokenizador robusto y la definición de etiquetas bien delimitadas contribuyeron a garantizar la coherencia del conjunto anotado.

Ejemplo de anotación La Tabla 3.3 muestra un ejemplo ilustrativo del formato de anotación.

archivo	línea	hablante	texto_completo	token	etiqueta
file_01.txt	12	speaker_1	ah sí. de las más veia	ah	es
file_01.txt	12	speaker_1	ah sí. de las más veia	sí	es
file_01.txt	12	speaker_1	ah sí. de las más veia	.	other
file_01.txt	12	speaker_1	ah sí. de las más veia	de	es
file_01.txt	12	speaker_1	ah sí. de las más veia	las	es
file_01.txt	12	speaker_1	ah sí. de las más veia	más	es
file_01.txt	12	speaker_1	ah sí. de las más veia	veia	mix
file_02.txt	34	speaker_5	se me incendió la estufa	se	es
file_02.txt	34	speaker_5	se me incendió la estufa	me	es
file_02.txt	34	speaker_5	se me incendió la estufa	incendió	es
file_02.txt	34	speaker_5	se me incendió la estufa	la	es
file_02.txt	34	speaker_5	se me incendió la estufa	estufa	es

Tabla 3.3: Ejemplo del formato de anotación manual en CSV.

3.5. Estadísticas descriptivas

Como resultado del proceso de anotación manual, se etiquetaron más de **30 000 tokens** del corpus, lo que permitió contar con un recurso suficientemente amplio para llevar a cabo los experimentos de clasificación. En total, se obtuvieron **31 054 tokens anotados**, cuya distribución por clase se muestra en la Tabla 3.4.

Etiqueta	Tokens	%
<i>es</i>	8.908	28,69 %
<i>pt</i>	14.313	46,09 %
<i>mix</i>	466	1,50 %
<i>ne</i>	257	0,83 %
<i>foreign</i>	17	0,05 %
<i>other</i>	7.093	22,84 %
Total	31.054	100 %

Tabla 3.4: Distribución de tokens por etiqueta en el corpus anotado manualmente.

Como puede observarse, el corpus está fuertemente dominado por las clases *pt* (46,09%) y *es* (28,69%), mientras que categorías como *mix* (1,50%), *ne* (0,83%) y *foreign* (0,05%) presentan una frecuencia mucho menor. La clase *other* constituye un 22,84% del total, lo que refleja la abundancia de signos, símbolos y elementos no léxicos propios de la transcripción oral.

Este desbalance es característico de los corpus de alternancia de códigos: las lenguas principales concentran la mayor cantidad de datos, en tanto que las clases minoritarias son más difíciles de capturar. En el contexto del aprendizaje automático, esta distribución desigual tiende a sesgar los modelos hacia las clases mayoritarias, lo que afecta la capacidad de reconocer correctamente los casos menos frecuentes, como los tokens *mix* o *foreign*.

Por este motivo, la **evaluación del desempeño** no puede basarse únicamente en la *accuracy*, ya que esta métrica resulta poco informativa cuando las clases están desbalanceadas. En su lugar, se prioriza el uso de métricas más expresivas como la **F1-score macro**, que combina la precisión (*precision*) y la exhaustividad (*recall*) para ofrecer una medida equilibrada del rendimiento por clase, ponderando a todas las etiquetas por igual. De esta manera, se obtiene una representación más justa del desempeño general del modelo, independientemente de la frecuencia de cada categoría.

3.6. Particionado en *train/dev/test*

Para la etapa experimental, el corpus anotado se dividió en tres subconjuntos disjuntos: *train*, *dev* y *test*. El objetivo fue disponer de conjuntos independientes para el entrenamiento, la validación de hiperparámetros y la evaluación final de los modelos, evitando la contaminación de datos entre ellos (*data leakage*).

El proceso de particionado se realizó **después de la anotación manual**, de modo que cada token ya contaba con su etiqueta correspondiente. Esto permitió garantizar que los subconjuntos conservaran la proporción relativa de clases y que ningún fragmento del mismo contexto conversacional apareciera en más de un conjunto. El muestreo se efectuó de forma estratificada por etiqueta, preservando la distribución global del corpus.

En total, el conjunto de **31.054 tokens anotados** se dividió de la siguiente manera: **16.255 tokens** para entrenamiento (52,3%), **7.179 tokens** para desarrollo (23,1%) y **7.620 tokens** para prueba (24,5%). La Tabla 3.5 muestra la distribución de etiquetas en cada subconjunto.

Etiqueta	Train	Dev	Test	Total
<i>es</i>	4.387	2.154	2.367	8.908
<i>pt</i>	7.706	3.197	3.410	14.313
<i>mix</i>	250	113	103	466
<i>ne</i>	127	84	46	257
<i>foreign</i>	10	3	4	17
<i>other</i>	3.775	1.628	1.690	7.093
Total	16.255	7.179	7.620	31.054

Tabla 3.5: Distribución de tokens por etiqueta en los subconjuntos de entrenamiento, desarrollo y prueba.

Como se observa, la proporción de clases se mantiene consistente entre los tres conjuntos, con ligeras variaciones esperables en las categorías menos frecuentes. Esto asegura que los modelos no se vean expuestos a un sesgo de distribución durante el entrenamiento y que las métricas de evaluación reflejen el comportamiento real del corpus.

La Figura 3.1 resume gráficamente el particionado general:

Train 16.255 (52,3 %)	Dev 7.179 (23,1 %)	Test 7.620 (24,5 %)
---------------------------------	------------------------------	-------------------------------

Figura 3.1: Particionado del corpus en conjuntos de entrenamiento, desarrollo y prueba.

Este esquema de división permite disponer de un conjunto de entrenamiento suficientemente amplio para el ajuste de parámetros, junto con subconjuntos independientes para la validación y la evaluación final, garantizando la reproducibilidad y comparabilidad de los resultados experimentales.

Capítulo 4

Experimentos

Este capítulo presenta los experimentos realizados sobre el corpus anotado de portuñol, con el objetivo de evaluar diferentes estrategias de **detección automática de idioma a nivel de token** en un contexto de *code-switching*. Los experimentos buscan analizar el desempeño de modelos que varían en complejidad y requisitos de datos, abarcando desde enfoques simples basados en reglas hasta arquitecturas neuronales modernas.

Se implementaron y compararon tres aproximaciones principales:

1. **Modelo heurístico basado en diccionarios:** un sistema de referencia construido a partir de listas de palabras y reglas explícitas para identificar tokens en español, portugués o mixtos.
2. **Modelo estadístico Naive Bayes:** un clasificador supervisado que utiliza información probabilística sobre las características ortográficas y de frecuencia de los tokens, evaluando distintas configuraciones del parámetro de suavización m .
3. **Modelos neuronales basados en BERT:** variantes multilingües y monolingües finamente ajustadas (*fine-tuned*) sobre el corpus de portuñol, con el fin de explorar la capacidad de las representaciones contextuales profundas para distinguir entre lenguas en contacto.

Cada sección describe las decisiones de implementación, los hiperparámetros empleados, las métricas utilizadas y los resultados obtenidos. Finalmente, se presenta una comparación conjunta de los tres enfoques, analizando sus fortalezas y limitaciones en función de la complejidad del modelo, el tamaño del corpus y la naturaleza del fenómeno lingüístico abordado.

4.1. Heurísticas basadas en reglas

Como primer enfoque se implementaron reglas heurísticas apoyadas en diccionarios de español, portugués e inglés, así como en criterios ortográficos y de contexto. El objetivo fue establecer una línea de base simple e **interpretable**, en el sentido de que sus decisiones pueden ser comprendidas y analizadas por un humano, siguiendo la distinción propuesta por [Clinciu y Hastie \(2019\)](#) entre transparencia, interpretabilidad y explicabilidad en los sistemas de inteligencia artificial.

4.1.1. Descripción del enfoque

El método asigna a cada *token* una etiqueta lingüística a partir de reglas determinísticas, contrastando con diccionarios de español, portugués e inglés y considerando ciertos patrones simples. Los diccionarios utilizados provienen de los recursos léxicos de `Hunspell`¹, que ofrecen amplia cobertura ortográfica y se emplean habitualmente en tareas de corrección y detección de idioma. Esta elección permitió contar con fuentes normativas estandarizadas para ambas lenguas, sin necesidad de construir manualmente listas léxicas.

La lógica de clasificación puede resumirse en los siguientes pasos, ilustrados en la Figura 4.1:

1. Coincidencia en un solo diccionario:

- Si el token aparece únicamente en el diccionario de español $\Rightarrow$ **es**.
- Si aparece únicamente en el diccionario de portugués $\Rightarrow$ **pt**.

2. Coincidencia en español y portugués:

En este caso de ambigüedad, se consulta la etiqueta asignada al token anterior (`prev_lang`). Si existe una etiqueta previa, se hereda; si no hay contexto anterior, se clasifica como **pt**.

3. Coincidencia en el diccionario de inglés:

El token se clasifica como **foreign**, ya que en el corpus las únicas ocurrencias de esta clase provenían del inglés.

4. Símbolos y marcadores:

Si el token pertenece al conjunto `{np, xx, xxx}`, o bien consiste únicamente en signos o puntuación, se etiqueta como **other**. Estas formas corresponden a marcas de anonimización y ruido introducidas en la transcripción original.

5. Entidades con mayúscula inicial:

Si el token comienza con mayúscula inicial en contexto, se clasifica como **ne**. Esta heurística se basa en la observación de que los nombres propios aparecen frecuentemente con esa forma gráfica.

¹`Hunspell` es un sistema de corrección ortográfica y segmentación morfológica de código abierto, ampliamente utilizado en procesadores de texto como LibreOffice, Firefox y Chrome. Permite acceder a listas léxicas y reglas morfológicas para más de 100 idiomas.

6. **Casos residuales o forzados:** Si el token no cumple ninguna de las reglas anteriores, se clasifica como **mix**. Además, se incluyó una lista **FORZAR_MIX** = {**home**, **homes**, **qua**} para asegurar que ciertos híbridos frecuentes se etiqueten directamente como **mix**.

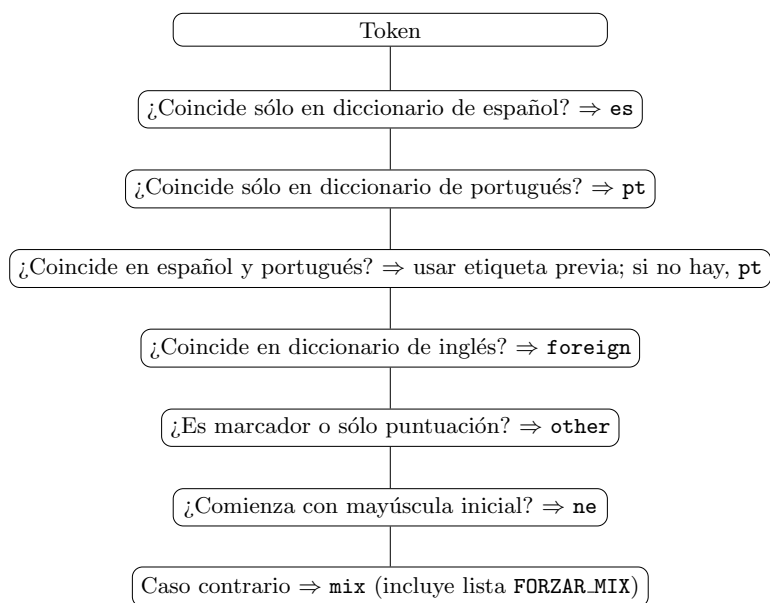


Figura 4.1: Flujo de reglas del clasificador heurístico basado en diccionarios.

Este enfoque ofrece una línea de base sólida e interpretable, aunque presenta limitaciones: (i) depende fuertemente de los diccionarios normativos, que no contemplan variación dialectal ni errores ortográficos; (ii) la clase **mix** funciona como contenedor residual de casos difíciles de modelar; y (iii) la desambiguación entre **es** y **pt** mediante el contexto inmediato es un criterio útil pero poco robusto en secuencias largas.

4.1.2. Resultados

La Tabla 4.1 presenta el desempeño del modelo heurístico en los conjuntos de desarrollo (*dev*) y prueba (*test*). En ambos casos se obtiene una exactitud global cercana al 88–89 %, impulsada por el alto rendimiento en las clases mayoritarias (*es*, *pt* y *other*). Sin embargo, el macro-F1 resulta sensiblemente menor (0,65 en *dev*, 0,64 en *test*), reflejando las dificultades para las clases minoritarias. El hecho de que los valores en *dev* y *test* sean tan cercanos evidencia una buena consistencia del enfoque y ausencia de sobreajuste.

Class	Precision	Recall	F1	Support
Conjunto <i>dev</i> (7.179 tokens)				
es	0,88	0,81	0,84	2.154
pt	0,88	0,90	0,89	3.197
mix	0,36	0,71	0,47	113
ne	0,65	0,77	0,71	84
foreign	0,00	0,00	0,00	3
other	1,00	0,98	0,99	1.628
Macro-F1		0,65		
Conjunto <i>test</i> (7.620 tokens)				
es	0,88	0,81	0,84	2.367
pt	0,88	0,89	0,89	3.410
mix	0,29	0,69	0,40	103
ne	0,63	0,89	0,74	46
foreign	0,00	0,00	0,00	4
other	1,00	0,98	0,99	1.690
Macro-F1		0,64		

Tabla 4.1: Resultados del clasificador heurístico en los conjuntos de desarrollo y prueba. La columna *Support* indica la cantidad de ejemplos de cada clase en el conjunto correspondiente.

La Figura 4.2 muestra la matriz de confusión, donde se observa que el modelo acierta con alta precisión en las clases mayoritarias, mientras que las categorías *mix* y *foreign* presentan errores más frecuentes. En particular, *mix* tiende a confundirse con *es* y *pt*, lo que confirma la dificultad de capturar tokens híbridos mediante reglas fijas. Las entidades nombradas (*ne*) alcanzan un *recall* elevado gracias a la heurística basada en mayúsculas iniciales, aunque con cierta pérdida de precisión por falsos positivos.

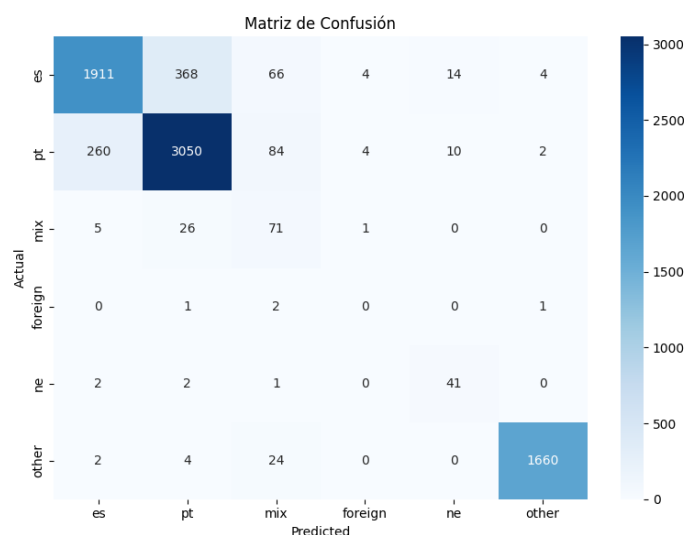


Figura 4.2: Matriz de confusión del clasificador heurístico en el conjunto de prueba.

En síntesis, este modelo heurístico constituye una línea de base sólida para las clases mayoritarias, pero insuficiente para categorías infrecuentes o ambiguas. Sus resultados motivan la exploración de enfoques más flexibles, como el modelo probabilístico Naive Bayes y los modelos neuronales basados en BERT, que se analizan en las siguientes secciones.

4.2. Naive Bayes

4.2.1. Descripción del enfoque

Como segundo enfoque se empleó un clasificador *Naive Bayes* a nivel de *token*. En particular, se utilizó **Complement Naive Bayes** (`ComplementNB`) (Rennie, Shih, Teevan, y Karger, 2003), una variante recomendada para clasificación de texto con *clases desbalanceadas*. La implementación proviene de `scikit-learn` (Pedregosa y cols., 2011)².

La representación se basó en ***n*-gramas de caracteres**, estándar en identificación de lenguas por su robustez ante tokens cortos o ruidosos (Jauhiainen y cols., 2019). En esta tesis, el *rango de n-gramas* (a, b) indica que se extraen todos los *n*-gramas de caracteres de longitudes $n \in [a, b]$ (inclusive). (La definición general de *n*-grama y la distinción entre *n*-gramas de palabra y de carácter se presentan en el Capítulo 2.)

4.2.2. Configuración y validación

El modelo se entrenó sobre *train* y la selección de hiperparámetros se realizó en *dev*. Se exploraron los siguientes valores:

- Suavizado de Laplace: $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0,5, 1, 2\}$.
- Rangos de *n*-gramas de caracteres: $(1, 5), (1, 3), (2, 5), (2, 6), (3, 5)$.

Luego de evaluar las combinaciones anteriores, la configuración que obtuvo el mejor rendimiento sobre *dev* fue: $\alpha = 10^{-4}$ y `n_gram_range = (1, 5)`, alcanzando un `Macro-F1 = 0,660`. Este resultado es coherente con la intuición de que incluir *n*-gramas cortos y medianos de carácter favorece la captura de patrones ortográficos útiles para *LID* a nivel de *token*.

²Documentación de `ComplementNB` en `scikit-learn`: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.ComplementNB.html

4.2.3. Resultados

La Tabla 4.2 reporta métricas por clase en *dev* y *test*. Se observa buen rendimiento en las clases mayoritarias y una mejora clara en *mix* respecto del enfoque heurístico. La generalización es estable: la *accuracy* es de 0,86 en *dev* y 0,85 en *test*, y el *macro-F1* se mantiene muy cercano entre ambos conjuntos.

Class	Precision	Recall	F1	Support
Conjunto <i>dev</i> (7.179 tokens)				
es	0,83	0,71	0,77	2.154
pt	0,82	0,90	0,86	3.197
mix	0,65	0,64	0,65	113
ne	0,71	0,68	0,70	84
foreign	0,00	0,00	0,00	3
other	1,00	0,99	0,99	1.628
Macro-F1		0,66		
Conjunto <i>test</i> (7.620 tokens)				
es	0,82	0,72	0,77	2.367
pt	0,82	0,89	0,85	3.410
mix	0,56	0,53	0,54	103
ne	0,61	0,78	0,69	46
foreign	0,00	0,00	0,00	4
other	1,00	1,00	1,00	1.690
Macro-F1		0,64		

Tabla 4.2: Resultados del clasificador **ComplementNB** en los conjuntos de desarrollo y prueba. La columna *Soporte* indica la cantidad de ejemplos por clase.

La matriz de confusión (Figura 4.3) muestra con mayor detalle los aciertos y errores del modelo. Se confirma el buen desempeño en *es*, *pt* y *other*, con la mayoría de los tokens correctamente identificados. No obstante, persisten confusiones entre *es* y *pt*, coherentes con el solapamiento léxico entre ambas lenguas. La clase *mix* es reconocida parcialmente, pero todavía se dispersa en *es* y *pt*, lo que evidencia la dificultad de capturar tokens híbridos. Las clases minoritarias, en particular *foreign*, apenas son reconocidas debido a su escaso soporte en el corpus (el F1 para esa clase es 0).

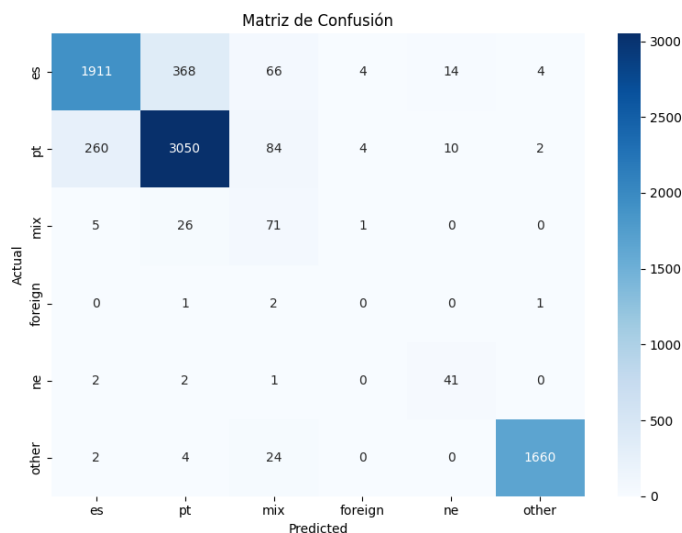


Figura 4.3: Matriz de confusión del clasificador `ComplementNB` en el conjunto de prueba.

En síntesis, `ComplementNB` mantiene buen desempeño en *es*, *pt* y *other* ($F1$ de 0,77, 0,85 y 1,00 en *test*) y mejora de forma clara la clase *mix* ($F1 = 0,54$ en *test*, frente a 0,40 del heurístico). Las *named entities* (*ne*) alcanzan $F1 = 0,69$, inferior al heurístico basado en mayúsculas iniciales, y *foreign* no muestra mejoras ($F1 = 0$). Globalmente, el *macro-F1* de Naive Bayes (0,64) empata al heurístico en *test*, con ventajas en *mix* y pequeñas pérdidas en *es/pt*. Estos resultados motivan la exploración de un modelo contextual con mayor capacidad de desambiguación, presentada en la Sección 4.3.

4.3. BERT

4.3.1. Descripción del enfoque

BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin y cols., 2019) es un modelo basado en la arquitectura *Transformer* (Vaswani y cols., 2017), que aprende representaciones contextuales de palabras a partir de grandes corpus monolingües o multilingües. En este trabajo se lo aplica a la tarea de *token classification*, donde cada token del corpus de portugués debe recibir una etiqueta de idioma (*es*, *pt*, *mix*, *ne*, *foreign*, *other*). La arquitectura consiste en añadir una capa de clasificación lineal sobre las salidas de BERT y ajustar el modelo mediante *fine-tuning* con los datos anotados manualmente.

4.3.2. Selección del modelo BERT

Antes del entrenamiento, se evaluó la calidad de la tokenización sobre un subconjunto de tokens etiquetados como *mix*, con el fin de analizar el comportamiento de los tokenizadores frente a formas híbridas y no estándar. Se compararon cinco variantes disponibles en la plataforma *Hugging Face*, todas preentrenadas y de acceso público:

- **bert-base-multilingual-cased** (mBERT)³: entrenado en más de 100 lenguas a partir de Wikipedia, incluyendo español, portugués y gallego (Devlin y cols., 2019).⁴ Su cobertura amplia y balanceada lo convierte en un modelo especialmente adecuado para tareas de identificación de idioma en contextos multilingües o de contacto entre lenguas romances.
- **bert-base-multilingual-uncased**⁵: versión no sensible a mayúsculas del modelo multilingüe.
- **dccuchile/bert-base-spanish-wwm-cased** (BETO)⁶: modelo monolingüe entrenado exclusivamente en español, con enmascaramiento de palabras completas (*Whole Word Masking*) (Cañete y cols., 2020).
- **neuralmind/bert-base-portuguese-cased** (BERTimbau)⁷: modelo monolingüe para portugués brasileño, preentrenado sobre el corpus *BrWaC* (Souza, Nogueira, y Lotufo, 2020).
- **dvilares/bertinho-gl-small-cased** (BERTinho)⁸: modelo entrenado en gallego, una lengua estrechamente relacionada con el portugués (Vilares, García, y Gómez-Rodríguez, 2021).

Esta comparación permitió identificar las diferencias de cobertura y segmentación entre modelos monolingües y multilingües, aspecto crucial para la representación de tokens híbridos en el portuñol. La selección final de modelos a utilizar en los experimentos se basó tanto en la cobertura léxica observada sobre el corpus como en el desempeño alcanzado tras el *fine-tuning*, detallado en las siguientes secciones.

³<https://huggingface.co/bert-base-multilingual-cased>

⁴De acuerdo con la documentación oficial de Google Research, el modelo multilingüe abarca las 100 Wikipedias más extensas, entre ellas la del gallego: <https://github.com/google-research/bert/blob/master/multilingual.md>.

⁵<https://huggingface.co/bert-base-multilingual-uncased>

⁶<https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased>

⁷<https://huggingface.co/neuralmind/bert-base-portuguese-cased>

⁸<https://huggingface.co/dvilares/bertinho-gl-small-cased>

4.3.3. Análisis de la tokenización

Antes de proceder al entrenamiento, se realizó un análisis comparativo de los tokenizadores de los modelos BERT considerados, con el fin de evaluar su capacidad para segmentar adecuadamente los *tokens mix*. Estos tokens híbridos, que combinan rasgos ortográficos y morfológicos del español y el portugués, tienden a fragmentarse en múltiples subtokens durante la tokenización, debido a que en su mayoría no pueden encontrarse en el vocabulario estándar de ninguno de los dos idiomas. Esta fragmentación puede afectar la coherencia de las representaciones aprendidas y, en consecuencia, el rendimiento del modelo.

El experimento tuvo como objetivo identificar qué modelo produce una segmentación más estable y compacta sobre palabras mixtas, es decir, cuál logra la mejor *cobertura subléxica*. Para ello se utilizaron las métricas propuestas por Rust y cols. (2021):

- **Subword Fertility**: número promedio de subtokens generados por cada palabra original.
- **Subword Proportion**: proporción de palabras que se dividen en más de un subtoken.

Ambas métricas se interpretan de manera inversa: valores más bajos indican menor fragmentación y, por tanto, una representación más eficiente del léxico.

El cálculo se realizó sobre el conjunto de 236 *tokens mix* distintos presentes en todo el corpus, considerando cada forma léxica única una sola vez. Este análisis se llevó a cabo utilizando los tokenizadores preentrenados de cada modelo, sin realizar ningún tipo de entrenamiento ni ajuste adicional, con el objetivo de evaluar exclusivamente su comportamiento intrínseco frente a formas híbridas. Los resultados se presentan en la Tabla 4.3.

Modelo	Subword Fertility	Subword Proportion (%)
mBERT-cased	2,25	88,56
mBERT-uncased	2,40	94,07
BERTinho (gl)	2,46	91,10
BETO (es)	2,42	91,53
BERTimbau (pt)	2,46	94,07

Tabla 4.3: Métricas de segmentación subléxica sobre tokens *mix* para los modelos evaluados. Valores menores indican menor fragmentación y, por tanto, mayor cobertura léxica.

Los resultados muestran que todos los modelos presentan un alto grado de fragmentación sobre tokens híbridos, con proporciones de división superiores al 88 %. El modelo **mBERT-cased** exhibe los mejores valores generales, con la menor *Subword Fertility* (2,25) y la proporción más baja de fragmentación (88,6 %), lo que sugiere que su vocabulario multilingüe logra representar con

mayor compacidad los rasgos combinados del español y el portugués. Por el contrario, **mBERT-uncased** y **BERTimbau** muestran mayor tendencia a la sobresegmentación (*Subword Proportion* del 94 %), lo que podría introducir ruido en los límites de palabra. **BETO** (español) y **BERTinho** (gallego) alcanzan valores intermedios, con coberturas similares (alrededor del 91 %), pero sin superar al modelo multilingüe.

En función de estos resultados, se seleccionaron los modelos **mBERT-cased** y **BERTinho** para realizar el *fine-tuning*. Ambos presentan las mejores condiciones para capturar fenómenos de mezcla entre lenguas romances: mBERT por su cobertura simultánea de español, portugués y gallego, y BERTinho por su entrenamiento en una lengua geográficamente y tipológicamente afin al portuñol.

4.3.4. Configuración y validación

Se realizaron experimentos de *fine-tuning* con dos modelos preentrenados: **mBERT-cased** y **BERTinho-gl-small-cased**. La selección de estos modelos se basó en los resultados del análisis de tokenización presentado en la sección anterior, donde mostraron las mejores métricas de *Subword Fertility* y *Subword Proportion*. Ambos presentan, además, características lingüísticas relevantes para el fenómeno del portuñol: mBERT por su cobertura multilingüe (incluyendo español, portugués y gallego) y BERTinho por su proximidad tipológica y geográfica al contexto de contacto entre español y portugués en el norte de Uruguay.

Cada modelo se entrenó sobre el conjunto *train* y se validó en *dev*. Se aplicó el criterio de **early stopping** para prevenir el sobreajuste, deteniendo el entrenamiento cuando la métrica de validación dejaba de mejorar durante varias épocas consecutivas.

La métrica de selección fue el **Macro-F1**, más adecuada que la exactitud (*accuracy*) en contextos de clases desbalanceadas. Durante la evaluación se reportaron también precisión, *recall* y F1 por clase, a fin de analizar con mayor detalle el desempeño de cada modelo en las diferentes etiquetas lingüísticas.

4.3.5. Resultados

Esta sección presenta los resultados obtenidos tras el *fine-tuning* de los modelos **mBERT** y **BERTinho** sobre el corpus anotado de portuñol. Ambos modelos fueron seleccionados según el análisis de segmentación subléxica descrito en la Sección 4.3.3, por su mayor capacidad de capturar transiciones entre español y portugués en tokens híbridos. El desempeño se evaluó en los conjuntos de desarrollo (*dev*) y prueba (*test*) empleando las métricas de precisión, *recall* y F1 por clase, además de la métrica global **Macro-F1**.

mBERT

La Tabla 4.4 muestra los resultados del modelo multilingüe **bert-base-multilingual-cased**. El modelo alcanza una exactitud global del 87 % en *dev* y del 86 % en *test*. El

Macro-F1 es de 0,77 y 0,67 respectivamente, evidenciando una ligera caída en generalización, especialmente en las clases minoritarias.

Class	Precision	Recall	F1	Support
Conjunto <i>dev</i> (7.179 tokens)				
es	0,85	0,73	0,79	2.154
pt	0,83	0,91	0,87	3.197
mix	0,61	0,69	0,65	113
ne	0,82	0,77	0,80	84
foreign	1,00	0,33	0,50	3
other	1,00	0,99	0,99	1.628
Macro-F1		0,77		
Conjunto <i>test</i> (7.620 tokens)				
es	0,84	0,73	0,78	2.367
pt	0,82	0,90	0,86	3.410
mix	0,61	0,59	0,60	103
ne	0,75	0,89	0,81	46
foreign	0,00	0,00	0,00	4
other	1,00	1,00	1,00	1.690
Macro-F1		0,67		

Tabla 4.4: Resultados del modelo mBERT en los conjuntos de desarrollo y prueba. La columna *Support* indica la cantidad de ejemplos de cada clase.

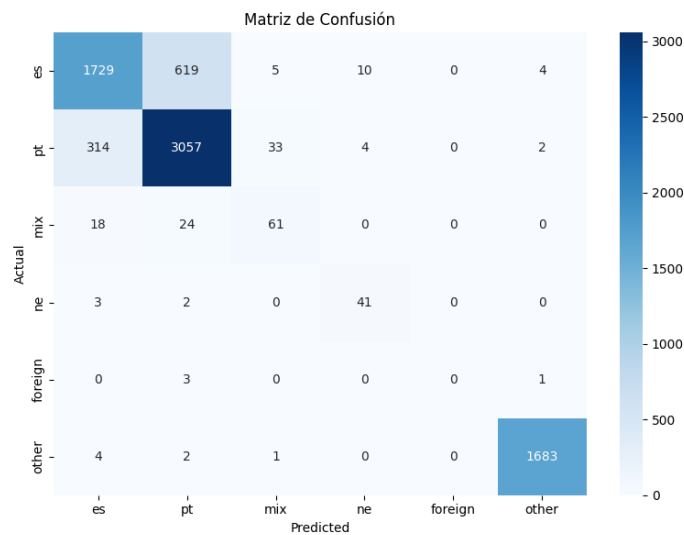


Figura 4.4: Matriz de confusión del modelo mBERT en el conjunto de prueba.

El modelo presenta un rendimiento sólido en las clases mayoritarias (*es*, *pt*, *other*), mientras que las clases menos representadas (*mix*, *foreign*, *ne*) continúan siendo más desafiantes. En particular, la clase *mix* muestra una mejora sustancial respecto de los enfoques heurísticos y Naive Bayes, confirmando que los modelos neuronales son más capaces de capturar regularidades ortográficas y contextuales en tokens híbridos. Sin embargo, se observa una caída de 0,10 puntos en el *Macro-F1* de *dev* a *test*, lo que sugiere un leve sobreajuste del modelo a los datos de entrenamiento.

BERTinho

La Tabla 4.5 muestra los resultados del modelo `dvilares/bertinho-gl-small-cased`, entrenado originalmente en gallego. Este modelo fue elegido por su proximidad lingüística con el portugués y su relevancia en contextos de contacto con el español.

Clase	Precisión	Recall	F1	Support
Conjunto <i>dev</i> (7,179 tokens)				
es	0,85	0,71	0,77	2.154
pt	0,82	0,91	0,86	3.197
mix	0,74	0,61	0,67	113
ne	0,78	0,80	0,79	84
foreign	0,00	0,00	0,00	3
other	1,00	0,99	0,99	1.628
Macro-F1		0,68		
Conjunto <i>test</i> (7,620 tokens)				
es	0,83	0,72	0,77	2.367
pt	0,82	0,90	0,86	3.410
mix	0,77	0,51	0,62	103
ne	0,67	0,72	0,69	46
foreign	0,00	0,00	0,00	4
other	1,00	1,00	1,00	1.690
Macro-F1		0,66		

Tabla 4.5: Resultados del modelo BERTinho en los conjuntos de desarrollo y prueba.

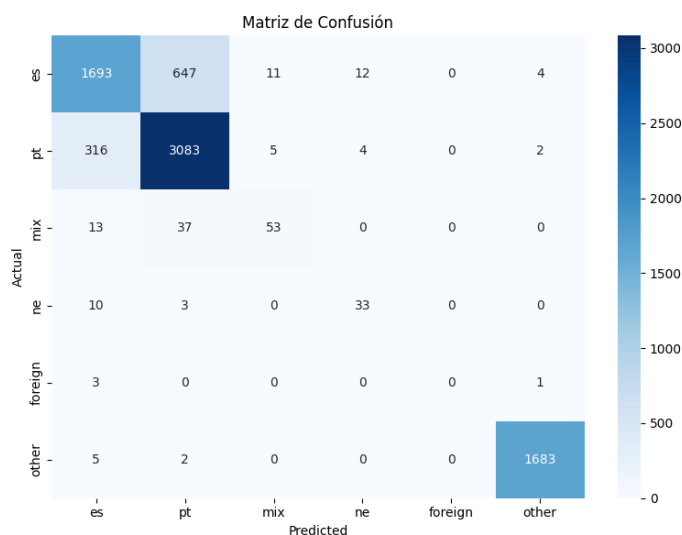


Figura 4.5: Matriz de confusión del modelo BERTinho sobre el conjunto de prueba.

En la Figura 4.5 se observa un patrón de comportamiento similar al de mBERT, con alta precisión en las clases *es*, *pt* y *other*. Sin embargo, se distingue una leve mejora en la identificación de tokens *mix*, que presentan menos confusiones hacia las etiquetas monolingües. Esto sugiere que el modelo entrenado en gallego logra capturar con mayor sensibilidad las transiciones morfológicas entre español y portugués, probablemente debido a la afinidad estructural del gallego con ambas lenguas. Pese a esta ventaja, las confusiones entre *es* y *pt* siguen siendo frecuentes, lo cual confirma el alto grado de solapamiento léxico entre las dos lenguas base del portuñol.

Análisis comparativo

Al comparar los dos modelos neuronales evaluados, mBERT y BERTinho, se observa un desempeño global muy similar, con una exactitud del 86 % en ambos casos. El modelo mBERT alcanza un *Macro-F1* ligeramente superior (0,67 frente a 0,66), mientras que BERTinho muestra una ventaja puntual en la clase *mix* ($F1 = 0,62$ frente a 0,60). Esta diferencia sugiere que el entrenamiento previo de BERTinho en gallego —una lengua que comparte rasgos estructurales y léxicos con el portugués, pero en contacto constante con el español— le permite capturar con mayor sensibilidad los patrones híbridos característicos del portuñol.

Ambos modelos exhiben una confusión recurrente entre *es* y *pt*, esperable dada la alta superposición léxica entre ambas lenguas. No obstante, BERTinho tiende a reducir parcialmente los falsos positivos de *mix* hacia *es* y *pt*, lo que se refleja en una mayor estabilidad en esa clase. Por su parte, mBERT mantiene

un comportamiento más equilibrado y generalizable, posiblemente debido a su entrenamiento multilingüe más amplio, que incluye al gallego entre sus lenguas de preentrenamiento.

En síntesis, los resultados de esta comparación evidencian que el conocimiento previo del gallego aporta beneficios concretos en la detección de tokens híbridos, aunque el modelo multilingüe continúa ofreciendo un desempeño global más estable y robusto.

4.4. Comparación de enfoques

En esta sección se contrastan los tres enfoques implementados: heurísticas basadas en reglas, Naive Bayes y BERT. El objetivo es analizar sus diferencias no solo en términos de desempeño global, sino también en la capacidad de cada modelo para manejar las clases minoritarias y los tokens híbridos, característicos del portugués.

4.4.1. Desempeño global

La Tabla 4.6 resume los resultados globales en los conjuntos de desarrollo (*dev*) y prueba (*test*), en términos de exactitud (*accuracy*) y Macro-F1.

Modelo	Desarrollo (dev)		Prueba (test)	
	Accuracy	Macro-F1	Accuracy	Macro-F1
Heurísticas basadas en reglas	0,89	0,65	0,88	0,64
Naive Bayes (ComplementNB)	0,86	0,66	0,85	0,64
BERT (mBERT fine-tuned)	0,87	0,77	0,86	0,67
BERTinho (fine-tuned)	0,86	0,68	0,86	0,66

Tabla 4.6: Comparación de desempeño global entre heurísticas, Naive Bayes y modelos BERT en los conjuntos de desarrollo y prueba.

Los resultados muestran que, aunque el enfoque heurístico alcanza la mayor *accuracy*, su *Macro-F1* permanece bajo debido al mal desempeño en las clases minoritarias. Naive Bayes mejora ligeramente la cobertura de *mix*, mientras que los modelos BERT —especialmente mBERT— logran un equilibrio más adecuado entre precisión y generalización, destacándose en escenarios con desbalance de clases.

4.4.2. Comparación por categoría

La Tabla 4.7 muestra los valores de F1 por clase en el conjunto de desarrollo y prueba para cada modelo.

Clase	Desarrollo (dev)				Prueba (test)			
	Heur.	NB	mBERT	BERTinho	Heur.	NB	mBERT	BERTinho
<i>es</i>	0,84	0,77	0,79	0,77	0,84	0,77	0,78	0,77
<i>pt</i>	0,89	0,86	0,87	0,86	0,89	0,85	0,86	0,86
<i>mix</i>	0,47	0,65	0,65	0,67	0,40	0,54	0,60	0,62
<i>ne</i>	0,71	0,70	0,80	0,79	0,74	0,69	0,81	0,69
<i>foreign</i>	0,00	0,00	0,50	0,00	0,00	0,00	0,00	0,00
<i>other</i>	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00
Macro-F1	0,65	0,66	0,77	0,68	0,64	0,64	0,67	0,66

Tabla 4.7: Comparación del desempeño por clase (F1-score) de los cuatro modelos en los conjuntos de desarrollo y prueba. Los mejores valores por fila se muestran en **negrita**.

En términos de desempeño por clase:

- **Clases mayoritarias (*es*, *pt*, *other*):** Todos los enfoques obtienen valores elevados de F1, superiores a 0,75. Las heurísticas muestran el mejor desempeño global en estas categorías, impulsadas por la coincidencia directa con los diccionarios normativos. Sin embargo, esta misma rigidez impide reconocer variantes no estándar del portugués, donde son frecuentes las adaptaciones fonéticas y ortográficas.
- **Tokens híbridos (*mix*):** Aquí se concentran las mayores diferencias entre enfoques. El modelo de **Naive Bayes** mejora de forma notable sobre las heurísticas (F1=0,54 vs. 0,40) al capturar regularidades ortográficas mediante n-gramas de caracteres. No obstante, los modelos BERT superan ampliamente ambos métodos: mBERT alcanza F1=0,60 y BERTinho F1=0,62. Este último parece beneficiarse del preentrenamiento en gallego, una lengua situada en una zona de contacto histórico entre el español y el portugués, lo que sugiere una transferencia positiva para la identificación de secuencias híbridas propias del portugués.
- **Entidades nombradas (*ne*):** Los modelos BERT logran el mejor rendimiento (F1=0,81 en mBERT, 0,69 en BERTinho), lo que refleja la ventaja del contexto bidireccional y la representación contextual frente a las reglas simples de capitalización. Estas diferencias muestran que, aunque el reconocimiento de nombres propios depende en parte de rasgos formales, el contexto circundante resulta determinante para evitar falsos positivos (por ejemplo, al inicio de oración).

- **Clases minoritarias (*foreign*):** Ninguno de los enfoques logra un rendimiento significativo debido al escaso número de ejemplos. Sin embargo, se observan algunos aciertos esporádicos en BERT, atribuibles al conocimiento previo adquirido durante el preentrenamiento multilingüe.

4.4.3. Discusión general

El análisis comparativo permite observar una progresión clara en la complejidad y cobertura de los modelos: las heurísticas ofrecen un punto de partida interpretable y rápido, Naive Bayes introduce generalización estadística a partir de regularidades ortográficas, y los modelos BERT incorporan sensibilidad contextual y robustez ante la variación lingüística.

Durante la etapa de anotación manual se identificaron varios patrones morfológicos recurrentes que ilustran los desafíos de clasificación. Por ejemplo: la palabra *fio* se utiliza para referirse a *hijo*, derivada de *filho* en portugués, pero en portugués estándar *fio* significa “hilo”. En ausencia de contexto, un modelo puramente léxico o estadístico no puede resolver esta ambigüedad, mientras que un modelo contextual como BERT dispone de información adicional para desambiguar según el entorno sintáctico y discursivo.

Asimismo, se observó un patrón fonético y ortográfico característico del portugués: la sustitución del dígrafo portugués *lh* por la vocal *i*, lo que da lugar a formas como:

$$\begin{aligned} \textit{melhor} \text{ (PT)} &\rightarrow \textit{meior} \rightarrow \textit{mejor} \text{ (ES)}, \\ \textit{ovelha} \text{ (PT)} &\rightarrow \textit{oveia} \rightarrow \textit{oveja} \text{ (ES)}, \\ \textit{trabalho} \text{ (PT)} &\rightarrow \textit{trabaio} \rightarrow \textit{trabajo} \text{ (ES)}. \end{aligned}$$

Estos fenómenos muestran la gradualidad de la mezcla lingüística y explican por qué los tokens híbridos (*mix*) son particularmente difíciles de clasificar: se ubican en un continuo entre ambas lenguas, sin pertenecer completamente a ninguna de ellas.

No obstante, el análisis de los resultados permite observar que los modelos BERT logran capturar parte de esta complejidad cuando cuentan con suficiente contexto. Por ejemplo, en el conjunto de prueba el modelo clasificó correctamente como *mix* el token *oveia* en el siguiente fragmento:

“ah, e agora, hoje no má tenho que levá água. tenho que chegar e botá água. eh. a muié e que vai inventar de botá as. as tábua esa na volta das planta dela, que as galinha e as **oveia** ’tão cumedo tudo”

De manera similar, se observó una clasificación correcta del token *fio* en el siguiente ejemplo del conjunto de prueba:

“desejo. desejo. de. de ter meus **fio** sempre todos junto”

En este caso, el contexto sintáctico y semántico asociado a la noción de familia y posesión (*meus*, *ter*, *todos junto*) favorece la interpretación de *fio* como una forma híbrida relacionada con *filho*, y no como el sustantivo portugués

estándar *fio* (“hilo”). Estos ejemplos ilustran cómo los modelos contextuales logran desambiguar tokens híbridos a partir de dependencias léxicas y discursivas más amplias, superando las limitaciones de enfoques puramente léxicos o estadísticos.

4.4.4. Reflexión final

Los resultados evidencian que el portuñol plantea desafíos únicos para el procesamiento automático del lenguaje: su naturaleza híbrida, su variabilidad ortográfica y su fuerte dependencia del contexto dificultan la tarea incluso para modelos avanzados. Sin embargo, el hecho de que los modelos neuronales —en especial BERT y BERTinho— logren mejoras claras en la detección de tokens *mix* marca un avance significativo hacia la construcción de herramientas adaptadas a variedades lingüísticas de frontera.

Más allá de los resultados numéricos, este estudio aporta evidencia empírica sobre cómo distintas aproximaciones computacionales capturan (o fallan en capturar) la dinámica del contacto lingüístico. De esta forma, sienta las bases para futuras investigaciones que integren enfoques híbridos entre lingüística y aprendizaje profundo, aprovechando la riqueza del portuñol como laboratorio natural para el estudio del *code-switching*.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

En este trabajo se abordó la tarea de *token classification* en portugués, con el objetivo de clasificar cada token en seis categorías: *es*, *pt*, *mix*, *ne*, *foreign* y *other*. Para ello se diseñó un corpus anotado manualmente y se exploraron tres enfoques principales: heurísticas basadas en reglas, Naive Bayes y BERT.

Los resultados obtenidos permiten extraer varias conclusiones:

- **Heurísticas:** constituyen una línea de base sólida para las clases mayoritarias, alcanzando una exactitud elevada (0,88) pero con limitaciones claras en categorías minoritarias (*mix*, *foreign*, *ne*), lo que se refleja en un Macro-F1 de apenas 0,64.
- **Naive Bayes:** ofrece una mejora relativa en la detección de *mix* respecto a las reglas, pero su desempeño global no supera a las heurísticas, manteniéndose en un Macro-F1 de 0,64 y con menor exactitud (0,85). Esto evidencia las dificultades de un modelo estadístico lineal frente a la complejidad del fenómeno híbrido.
- **BERT:** mostró el mejor equilibrio entre clases, con un Macro-F1 de 0,67 en el conjunto de prueba. Si bien su exactitud es similar a la de Naive Bayes (0,86), el modelo neuronal logró capturar mejor las categorías ambiguas, en particular *mix* y *ne*. Estos resultados confirman la utilidad de representaciones contextuales profundas en un escenario multilingüe y desbalanceado.

En términos generales, se observó que la exactitud global puede ser elevada incluso con enfoques simples, pero el verdadero desafío radica en equilibrar el desempeño entre clases. El modelo BERT, aun sin una gran cantidad de datos,

se mostró más robusto frente a esta dificultad, lo cual refuerza la pertinencia de emplear técnicas de aprendizaje profundo para abordar fenómenos de contacto lingüístico.

5.2. Trabajo futuro

Este estudio sienta las bases para la modelización computacional del portuñol y abre diversas líneas de trabajo a futuro que podrían fortalecer y ampliar sus resultados. En primer lugar, una dirección evidente es la ampliación y diversificación del corpus, incorporando más hablantes, regiones y contextos comunicativos. Esto permitiría no solo mejorar el balance de clases y la cobertura léxica, sino también capturar variación sociolingüística actualmente subrepresentada, lo que beneficiaría el entrenamiento de modelos más generales.

En paralelo, sería valioso avanzar hacia un esquema de etiquetado más fino y multidimensional. La distinción entre préstamos integrados y no integrados, o entre distintos tipos de entidades nombradas, posibilitaría un análisis más detallado de los fenómenos de contacto y contribuiría a una comprensión más precisa de las dinámicas de mezcla lingüística.

A partir de las limitaciones observadas en los enfoques puramente basados en reglas o en aprendizaje profundo, surge también la necesidad de explorar estrategias híbridas. Combinar reglas lingüísticas explícitas con representaciones neuronales permitiría aprovechar la interpretabilidad de las primeras y la capacidad de generalización de las segundas, potenciando así la robustez y explicabilidad de los modelos.

Otra línea de trabajo prometedora es el desarrollo de modelos especializados. La adaptación de arquitecturas recientes, como XLM-RoBERTa o modelos *domain-adapted* entrenados específicamente sobre datos de portuñol, podría mejorar la cobertura y reducir la fragmentación subléxica detectada en los experimentos de esta tesis (Liu y cols., 2019).

Finalmente, las aplicaciones prácticas de esta línea de investigación abren un campo de gran potencial. La detección automática de portuñol y de sus matices podría integrarse en herramientas de enseñanza bilingüe, en estudios de variación lingüística asistidos por computadora, o en sistemas de comunicación adaptados a comunidades fronterizas. De este modo, los avances en procesamiento del lenguaje natural podrían contribuir no solo al conocimiento académico, sino también al fortalecimiento y visibilización de las variedades híbridas del español y el portugués en el Cono Sur.

En suma, este trabajo constituye un primer paso hacia la representación computacional del portuñol, demostrando que las técnicas actuales de PLN pueden aplicarse con éxito a variedades en contacto. Al mismo tiempo, pone de relieve la importancia de seguir ampliando los recursos, optimizando los modelos y explorando enfoques integradores para alcanzar una caracterización más completa y sensible a la riqueza lingüística de esta frontera. Con el fin de fomentar la reproducibilidad y la transparencia, se prevé además publicar el código y

los scripts experimentales en un repositorio público de GitLab¹.

¹<https://gitlab.fing.edu.uy/rodrigo.souza/fronterianlp>

Referencias

- Aguilar, G., Kar, S., y Solorio, T. (2020). Lince: A centralized benchmark for linguistic code-switching evaluation. En *Proceedings of the 12th language resources and evaluation conference* (pp. 1803–1813).
- Brown, P. F., y cols. (2020). What works and what doesn't in language identification: A large-scale evaluation. En *Proceedings of the 12th language resources and evaluation conference (lrec 2020)* (pp. 1498–1506).
- Bullock, B. E., y Toribio, A. J. (2009). *The cambridge handbook of linguistic code-switching*. Cambridge University Press.
- Carvalho, A. M. (2007). Diagnóstico sociolingüístico de comunidades escolares fronterizas en el norte de uruguay. En C. Brovetto, J. Geymonat, y N. Brian (Eds.), *Portugués del uruguay y educación bilingüe* (pp. 49–98). Administración Nacional de Educación Pública, Uruguay.
- Carvalho, A. M. (2014). Introduction: Towards a sociolinguistics of the border. *International Journal of the Sociology of Language*, 227(1), 1–7. doi: 10.1515/ijsl-2013-0084
- Carvalho, A. M. (2016). The analysis of languages in contact: A case study through a variationist lens. *Cadernos de Estudos Linguísticos*, 58(3), 401–424. doi: 10.20396/cel.v58i3.8646468
- Carvalho, A. M. (2022). Contact between spanish and portuguese in south america. En *Oxford research encyclopedia of linguistics*. Oxford University Press. doi: 10.1093/acrefore/9780199384655.013.450
- Carvalho, A. M., y Christoffersen, K. (2022). On playful language divergences: Code-switching among spanish-portuguese bilinguals. *Semas*, 3(6), 31–49. doi: 10.18800/semas.202201.003
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., y Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. En *Pml4dc at iclr 2020*.
- Chiruzzo, L., Agüero-Torales, M. M., Giménez-Lugo, G., Alvarez, A., Rodríguez, Y., Góngora, S., y Solorio, T. (2023). Overview of gua-spa at iberlef 2023: Guarani-spanish code switching analysis. *Procesamiento del Lenguaje Natural*, 71, 321–328. Descargado de https://www.researchgate.net/publication/373649771_Overview_of_GUA-SPA_at_IberLEF_2023_Guarani-Spanish_Code_Switching_Analysis doi: 10.26342/2023-71-25

- Clinciu, M. A., y Hastie, H. F. (2019). A survey of explainable AI terminology. En *Proceedings of the 1st workshop on interactive natural language technology for explainable artificial intelligence (nl4xai)* (pp. 8–13). Tokyo, Japan: Association for Computational Linguistics. Descargado de <https://aclanthology.org/W19-8403/>
- Deka, P., y cols. (2022). Bert-based language identification in code-mix. En *Proceedings of the 19th international conference on natural language processing (icon-wlli)* (pp. 14–22). Descargado de <https://aclanthology.org/2022.icon-wlli.3>
- Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.
- Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.
- Garrido Sánchez-Andrade, B. (2025). *CLARU: Corpus Lingüístico de Artigas RurUrbano*. LaRS - Language Repository of Switzerland. Descargado de <https://doi.org/10.48656/2cs5-mv82> doi: 10.48656/2cs5-mv82
- Hughes, B., Baldwin, T., Bird, S., Nicholson, J., y MacKinlay, A. (2006). Reconsidering language identification for written language resources. En *Proceedings of the 5th international conference on language resources and evaluation (lrec)* (pp. 485–488).
- Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., y Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65, 675–782. doi: 10.1613/jair.1.11447
- Joshi, A. K. (1982). Processing of sentences with intra-sentential code-switching. En *Proceedings of coling 1982: The 9th international conference on computational linguistics* (pp. 145–150). Prague, Czechoslovakia: North-Holland Publishing Company.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., y Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 6282–6293. Descargado de <https://aclanthology.org/2020.acl-main.560>
- Jurafsky, D., y Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Draft Edition, Stanford University. Descargado de <https://web.stanford.edu/~jurafsky/slp3/>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lucas, A., Baladón, A., Pardiñas, V., Agüero-Torales, M., Góngora, S., y Chiruzzo, L. (2024). Grammar-based data augmentation for low-resource languages: The case of guarani–spanish neural machine translation. En *Proceedings of the 2024 conference of the north american chapter of the*

- association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 6385–6397). Association for Computational Linguistics. Descargado de <https://aclanthology.org/2024.naacl-long.354>
- Lui, M., y Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. *Proceedings of the ACL 2012 System Demonstrations*, 25–30.
- Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
- Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.
- Myers-Scotton, C. (1993). *Social motivations for codeswitching: Evidence from africa*. Oxford University Press.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Descargado de <https://jmlr.org/papers/v12/pedregosa11a.html>
- Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 18(7-8), 581–618.
- Rennie, J. D. M., Shih, L., Teevan, J., y Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. En *Proceedings of the 20th international conference on machine learning (icml-2003)* (pp. 616–623). Washington, DC, USA: AAAI Press. Descargado de <https://www.cs.cmu.edu/~enron/naivebayes.pdf>
- Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., y Gurevych, I. (2021). How good is your tokenizer? on the monolingual performance of multilingual language models. En *Proceedings of the 59th annual meeting of the association for computational linguistics (acl)*. Descargado de <https://arxiv.org/abs/2012.15613>
- Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., ... González, J. (2014). Overview for the first shared task on language identification in code-switched data. En *Proceedings of the first workshop on computational approaches to code switching* (pp. 62–72).
- Souza, F., Nogueira, R., y Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. En *9th brazilian conference on intelligent systems, BRACIS, rio grande do sul, brazil, october 20-23 (to appear)*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. En *Proceedings of the 31st international conference on neural information processing systems (neurips 2017)* (pp. 6000–6010). Curran Associates.
- Vilares, D., García, M., y Gómez-Rodríguez, C. (2021). Bertinho: Galician bert representations. En *Proceedings of the iberian languages evaluation forum (iberlef 2021)*. Spain: CEUR Workshop Proceedings. Descargado de <https://arxiv.org/abs/2103.13799>