Clasificación de Textos de Inspecciones

Grupo 20 - Laboratorio de Datos

Matías Morán, Gonzalo Ruarte, Alejandro Somoza

Julio 08, 2022

FCEyN, UBA

El problema a resolver

Inspecciones:

- Control de equipos: Se verifican equipos eléctricos mediante inspecciones dirigidas a domicilios.

El problema a resolver

Inspecciones:

- Control de equipos: Se verifican equipos eléctricos mediante inspecciones dirigidas a domicilios.
- Verificación de Estado: Las cuadrillas se dirigen al sitio para verificar el estado de una instalación.

El problema a resolver

La inspección culmina con un...

El problema a resolver

Un supervisor clasifica la inspección en una de las siguientes clases:

- Intrusión
- Reparación
- Marketing
- Visitada
- Sin realizar
- Seguridad
- Ok equipo
- Sin tarea

El problema a resolver

Especialistas verifican la clasificación y la aprueban o modifican según su criterio.

- Por semana revisan cierta cantidad de inspecciones.

El problema a resolver

Especialistas verifican la clasificación y la aprueban o modifican según su criterio.

- Por semana revisan cierta cantidad de inspecciones.
- Tarea que consume **tiempo** y **esfuerzo** pero importa para direccionar próximas inspecciones exitosas.

El problema a resolver

Especialistas verifican la clasificación y la aprueban o modifican según su criterio.

- Por semana revisan cierta cantidad de inspecciones.
- Tarea que consume **tiempo** y **esfuerzo** pero importa para direccionar próximas inspecciones exitosas.
- Hay un **costo** referido a la logística y a la deficiente asignación de la cuadrilla.

Solución propuesta

Formalmente es un problema de **clasificación supervisada multiclase**.

- Los textos clasificados y controlados por los expertos pueden ser utilizados para el entrenamiento de un **clasificador**.

Solución propuesta

Formalmente es un problema de **clasificación supervisada multiclase**.

- Los textos clasificados y controlados por los expertos pueden ser utilizados para el entrenamiento de un **clasificador**.
- El conjunto inicial de textos verificados por los especialistas forman la base de un **set de entrenamiento**.

Solución propuesta

Formalmente es un problema de **clasificación supervisada multiclase**.

- Los textos clasificados y controlados por los expertos pueden ser utilizados para el entrenamiento de un **clasificador**.
- El conjunto inicial de textos verificados por los especialistas forman la base de un **set de entrenamiento**.
- El clasificador puede **predecir** las etiquetas de **nuevos textos** correspondientes a inspecciones ulteriores.

Solución propuesta

Formalmente es un problema de **clasificación supervisada multiclase**.

- Los textos clasificados y controlados por los expertos pueden ser utilizados para el entrenamiento de un **clasificador**.
- El conjunto inicial de textos verificados por los especialistas forman la base de un **set de entrenamiento**.
- El clasificador puede **predecir** las etiquetas de **nuevos textos** correspondientes a inspecciones ulteriores.
- De esta forma se ahorra tiempo y se abaratan los costos.

Antecedentes

- Es posible encontrar ejemplos de aplicaciones similares como el análisis de sentimientos.

Antecedentes

- Es posible encontrar ejemplos de aplicaciones similares como el análisis de sentimientos.
- El clasificador de noticias de la agencia Reuters: la colección Reuters-21578

Antecedentes

- Es posible encontrar ejemplos de aplicaciones similares como el análisis de sentimientos.
- El clasificador de noticias de la agencia Reuters: la colección Reuters-21578
- El grupo de NLP de Stanford tiene evaluaciones sobre la misma accesibles en: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-text-classification-1.html

Antecedentes

- Es posible encontrar ejemplos de aplicaciones similares como el análisis de sentimientos.
- El clasificador de noticias de la agencia Reuters: la colección Reuters-21578
- El grupo de NLP de Stanford tiene evaluaciones sobre la misma accesibles en: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-text-classification-1.html
- Se emplean diferentes algoritmos para la clasificación de los textos.

Preprocesamiento del dataset

Una vez obtenido el conjunto de textos inicial, se empezaron a notar distintos tipos de problemas:

- Los textos son ingresados por operarios con diferente nivel educativo.

Preprocesamiento del dataset

Una vez obtenido el conjunto de textos inicial, se empezaron a notar distintos tipos de problemas:

- Los textos son ingresados por operarios con diferente nivel educativo.
- Existen errores de ortografía.

Preprocesamiento del dataset

Una vez obtenido el conjunto de textos inicial, se empezaron a notar distintos tipos de problemas:

- Los textos son ingresados por operarios con diferente nivel educativo.
- Existen errores de ortografía.
- Es común el uso de abreviaturas.

Preprocesamiento del dataset

Una vez obtenido el conjunto de textos inicial, se empezaron a notar distintos tipos de problemas:

- Los textos son ingresados por operarios con diferente nivel educativo.
- Existen errores de ortografía.
- Es común el uso de abreviaturas.
- Las clases no están balanceadas.

Preprocesamiento del dataset

Una vez obtenido el conjunto de textos inicial, se empezaron a notar distintos tipos de problemas:

- Los textos son ingresados por operarios con diferente nivel educativo.
- Existen errores de ortografía.
- Es común el uso de abreviaturas.
- Las clases no están balanceadas.
- Existen ambigüedades en las clasificaciones hechas por los especialistas.

Preprocesamiento del dataset

Si bien fue un proceso iterativo con muchas idas y vueltas, el mismo puede describirse de la siguiente forma:

- Se balancearon las clases teniendo en cuenta que **intrusión**, **reparación** y **seguridad** son las más importantes.

Preprocesamiento del dataset

Si bien fue un proceso iterativo con muchas idas y vueltas, el mismo puede describirse de la siguiente forma:

- Se balancearon las clases teniendo en cuenta que **intrusión**, **reparación** y **seguridad** son las más importantes.
- Se eliminaron acentos.

Preprocesamiento del dataset

Si bien fue un proceso iterativo con muchas idas y vueltas, el mismo puede describirse de la siguiente forma:

- Se balancearon las clases teniendo en cuenta que **intrusión**, **reparación** y **seguridad** son las más importantes.
- Se eliminaron acentos.
- Se pasaron los textos a minúscula.

Preprocesamiento del dataset

Si bien fue un proceso iterativo con muchas idas y vueltas, el mismo puede describirse de la siguiente forma:

- Se balancearon las clases teniendo en cuenta que **intrusión**, **reparación** y **seguridad** son las más importantes.
- Se eliminaron acentos.
- Se pasaron los textos a minúscula.
- Se estandarizaron abreviaturas.

Preprocesamiento del dataset

Si bien fue un proceso iterativo con muchas idas y vueltas, el mismo puede describirse de la siguiente forma:

- Se balancearon las clases teniendo en cuenta que **intrusión**, **reparación** y **seguridad** son las más importantes.
- Se eliminaron acentos.
- Se pasaron los textos a minúscula.
- Se estandarizaron abreviaturas.
- Se corrigieron errores de ortografía.

Preprocesamiento del dataset

Si bien fue un proceso iterativo con muchas idas y vueltas, el mismo puede describirse de la siguiente forma:

- Se balancearon las clases teniendo en cuenta que **intrusión**, **reparación** y **seguridad** son las más importantes.
- Se eliminaron acentos.
- Se pasaron los textos a minúscula.
- Se estandarizaron abreviaturas.
- Se corrigieron errores de ortografía.
- Se filtraron las stopwords (nombres, apellidos, monosílabos, conectores).

Preprocesamiento del dataset

Si bien fue un proceso iterativo con muchas idas y vueltas, el mismo puede describirse de la siguiente forma:

- Se balancearon las clases teniendo en cuenta que **intrusión**, **reparación** y **seguridad** son las más importantes.
- Se eliminaron acentos.
- Se pasaron los textos a minúscula.
- Se estandarizaron abreviaturas.
- Se corrigieron errores de ortografía.
- Se filtraron las stopwords (nombres, apellidos, monosílabos, conectores).
- (Los conectores “sin”, “no”, “con”, “ya” se mantienen ya que forman bigramas con palabras clave de la operatoria, como por ejemplo "sin fraude" vs "con fraude", que denota una intrusión o no en el equipo)

Preprocesamiento del dataset

Los pasos anteriores implicaron la utilización de **heurísticas** (muy similares a las vistas en clase, pero hay que tener en cuenta el dominio del problema) y diccionarios ad-hoc.

```
Termino_Abreviado,Termino_Completo
at,cambio medidor
-at,cambio medidor
a.t.,cambio medidor
a.t,cambio medidor
at.,cambio medidor
",at",cambio medidor
a/t,cambio medidor
a t,cambio medidor
r,intrusion
b,intrusion
m,intrusion
f,intrusion
t,normalizo obligaciones contractuales
si,seguridad
svp,seguridad
s.v.p.,seguridad
s.v.p,seguridad
s/a,ok_equipo
r - h,sin_tarea
n/l,sin_tarea
op,sin_tarea
san,ok_equipo
```

```
Termino_Input,Termino_Replace
anomalias,anomalia
baremos,baremo
cambia,cambio
cargas,carga
clandestino,clandestina
comunitaria,comunitario
concentrica,concentrico
conectada,conectado
conectadas,conectado
conectados,conectado
conecta,conecto
configura,configuro
conflictivos,conflictivo
constatada,constato
constata,constato
construyendo,construccion
consultorios,consultorio
consumos,consumo
contactos,contacto
convencionales,convencional
correspondientes,correspondiente
cortada,cortado
```

Figure 1: Ejemplos de diccionarios

Preprocesamiento del dataset

Los pasos anteriores implicaron la utilización de **heurísticas** (muy similares a las vistas en clase, pero hay que tener en cuenta el dominio del problema) y diccionarios ad-hoc.

```
Termino_Input,Termino_Replace
construccionvivienda,construccion vivienda
consumomovil,consumo movil
construccionmovil,construccion movil
conhidro,con hidro
consumose,consmo se
construccionobservaciones,construccion observaciones
conectarobservacion,conectar observacion
construccionsin,construccion sin
contruccionobservacion,construccion observacion
consumocuadrilla,consumo cuadrilla
contratistasin,contratista sin
concarga,con carga
contracurva,contra curva
conser,con ser
convencionalverifico,convencional verifico
consumocaja,consumo caja
construccionmedidor,construccion medidor
convencionalmolina,convencional molina
consumobrito,consumo brito
construccionse,construccion se
construccionulitizan,construccion utilizan
construccionnombre,construccion nombre
```

```
Termino_Input,Termino_Replace
abaj,abajo
abajos,abajo
abjo,abajo
yabajo,abajo
abandomada,abandonada
axion,accion
ccion,accion
acoemtida,acometida
acom,acometida
acomedida,acometida
acomet,acometida
acometidad,acometida
acometido,acometida
acometira,acometida
acomitida,acometida
acommetida,acometida
acomwtida,acometida
cometida,acometida
acometiday,acometida y
aconsejaretirarse,aconseja retirarse
adj,adjunta
```

Figure 1: Ejemplos de diccionarios

Preprocesamiento del dataset

Se verificó la clasificación base realizada por cada especialista sobre el dataset depurado y se observó que **distintos especialistas clasifican** a **textos similares** de **diferente** manera. Se realizó un **análisis de similaridad** para atacar este problema:

- Se armó una matriz de similaridad de textos mediante tfidf o lstm.

Preprocesamiento del dataset

Se verificó la clasificación base realizada por cada especialista sobre el dataset depurado y se observó que **distintos especialistas clasifican** a **textos similares** de **diferente** manera. Se realizó un **análisis de similaridad** para atacar este problema:

- Se armó una matriz de similaridad de textos mediante tfidf o lstm.
- En cada posición hay un valor de 0 a 1 que indica cuan similares son los textos.

Preprocesamiento del dataset

Se verificó la clasificación base realizada por cada especialista sobre el dataset depurado y se observó que **distintos especialistas clasifican** a **textos similares** de **diferente** manera. Se realizó un **análisis de similaridad** para atacar este problema:

- Se armó una matriz de similaridad de textos mediante tfidf o lstm.
- En cada posición hay un valor de 0 a 1 que indica cuan similares son los textos.
- Aquellos con un valor mayor a 0.75 y que hayan tenido clasificaciones ambiguas fueron mandados nuevamente a los especialistas para que se pongan de acuerdo.

Preprocesamiento del dataset

0.9	intr	medidor robado de otra cuenta	repar	se cambio medidor.
0.9	intr	medidor robado de otra cuenta	repar	se cambio medidor,tapa y termica trifasico
0.93	intr	medidor robado de otra cuenta	repar	se cambio medidor.golpeado
0.91	intr	medidor robado de otra cuenta	repar	medidor monofasico quemado.
0.91	intr	medidor robado de otra cuenta	repar	se cambio medidor trifasico quemado en bornera. se cambio tapa
0.9	intr	medidor robado de otra cuenta	repar	se verifico medidor instalacion y se procedio a cambiarlo
0.91	intr	medidor robado de otra cuenta	repar	cambio de medidor inspecciono instalacion se cambio frente de tapa antihurto
0.92	intr	medidor robado de otra cuenta	repar	verificacion completa , cambio de medidor
0.92	intr	medidor robado de otra cuenta	repar	se cambio medidor monof se cambio termica y tapa.
0.92	intr	medidor robado de otra cuenta	repar	se realiza cambio de medidor
0.92	intr	medidor robado de otra cuenta	repar	se cambio ramal trifasico y medidor
0.91	intr	medidor robado de otra cuenta	repar	se cambio medidor y tapa trifasica.
0.91	intr	medidor robado de otra cuenta	repar	se cambio medidor .termica y tapas linderas
0.92	intr	medidor robado de otra cuenta	repar	medidor con display apagado
0.9	intr	medidor robado de otra cuenta	repar	se cambio medidor.monof.cambio tapa y term.sello
0.91	intr	medidor robado de otra cuenta	repar	se cambia medidor y se instala caja antihurto
0.9	intr	medidas ya fue retirado. se llamo no se encuentra nadie en la casa.se retiro conexion directa desde linea	repar	inspeccion completa, dando mal el resultado, se cambia medidor trifasico y tapa rec/t.
0.9	intr	medidas ya fue retirado. se llamo no se encuentra nadie en la casa.se retiro conexion directa desde linea	repar	se verifico medidor con display apagado, se cambio trifasico y tapa.
0.9	intr	medidas ya fue retirado. se llamo no se encuentra nadie en la casa.se retiro conexion directa desde linea	repar	se coloco caja antihurto se cambio medidor quemado monofasico mas termica
0.91	intr	medidas ya fue retirado. se llamo no se encuentra nadie en la casa.se retiro conexion directa desde linea	repar	se llega al lugar y se charla con el cliente verificacndo el medidor el miamo se encontraba si
0.9	intr	medidas ya fue retirado. se llamo no se encuentra nadie en la casa.se retiro conexion directa desde linea	repar	se cambio medidor monofasico.es una obra no esta en uso
0.91	intr	medidas ya fue retirado. se llamo no se encuentra nadie en la casa.se retiro conexion directa desde linea	repar	se cambio medidor monofasico y se cambio tapa y termica.demoras para cerrar trabajo wo
0.9	intr	medidas ya fue retirado. se llamo no se encuentra nadie en la casa.se retiro conexion directa desde linea	repar	se cambia medidor y se instala caja antihurto
0.92	intr	se cambio medidor monofasico sin precinto de carcasa colocando caja antihurto con termica	repar	se cambio medidor,tapa y termica trifasico
0.93	intr	se cambio medidor monofasico sin precinto de carcasa colocando caja antihurto con termica	repar	verificacion completa cambio de medidor monofasico verificacion de funcionamiento
0.91	intr	se cambio medidor monofasico sin precinto de carcasa colocando caja antihurto con termica	repar	se cambik medidor monofasico por encontrarse obsoleto y se precinto tapa de bornera. se

Figure 2: Ejemplo de análisis de similaridad

Preprocesamiento del dataset

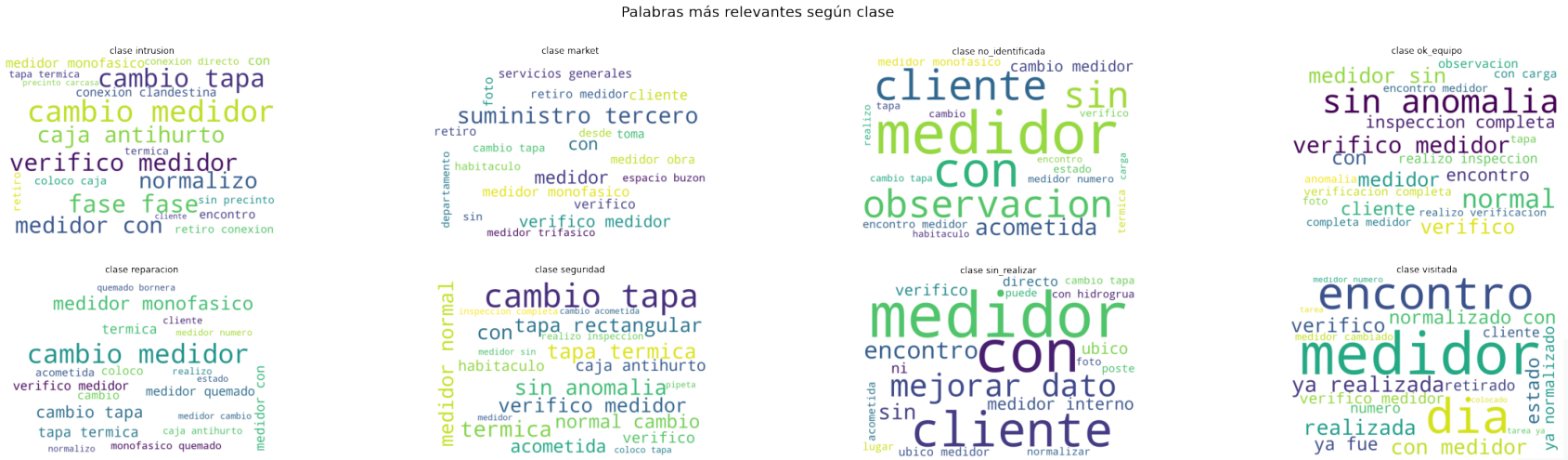


Figure 3: Wordclouds de cada clase luego del preprocesamiento.

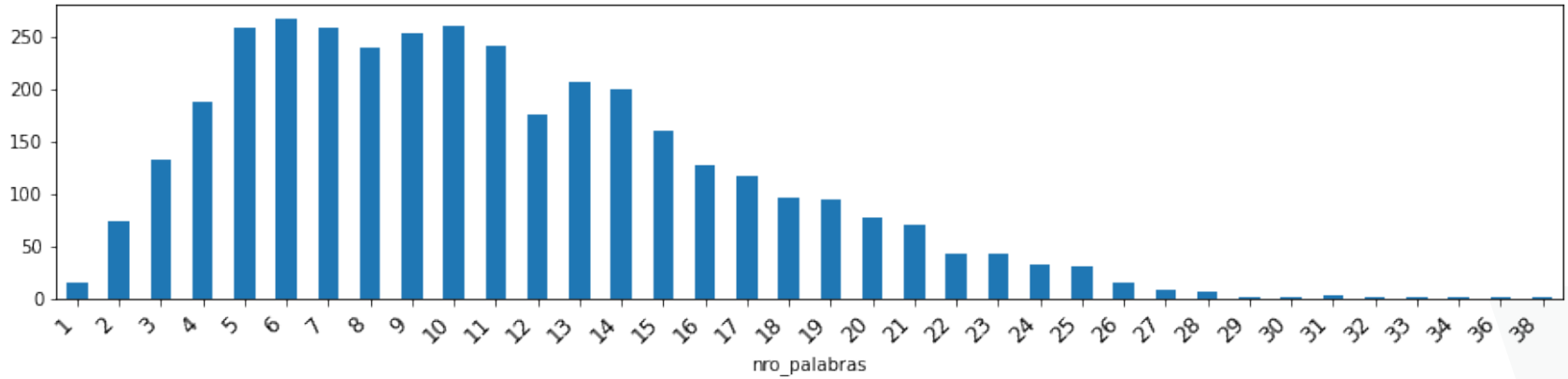
Preprocesamiento del dataset

texto_original	texto_etl
se normalizo medidor lindero XXXXXXX estado XXXXX/XXXXX cambiando tapa y acometida.	normalizo medidor lindero estado cambiando tapa acometida
se realizo verificacion completa se coloco caja antihurto se cambio pipeta rota energia activa: XXXXX energia reactiva: XXXXXX	realizo verificacion completa coloco caja antihurto cambio pipeta rota energia activa energia reactiva
se corto conexion directa sin medidor. en pje XXX camino Y XXXX	corto conexion directo sin medidor pje camino Y
se cambio medidor monofasico y normaliza.	cambio medidor monofasico normalizo
se retira conexion clandestina se normaliza suministro.	retira conexion clandestina normalizo suministro
numerador manipulado se cambia medidor.	numerador manipulado cambio medidor
cambio de medidor monofasico golpeado con visor roto y tapa monofasica obsoleta.	cambio medidor monofasico golpeado con visor roto tapa monofasica obsoleta
medidor manipulado fase y neutro invertidos la fase ingresaba por bornera de neutro y el neutro por bornera de fase se normalizo.	medidor manipulado fase neutro invertidos fase ingresaba bornera neutro neutro bornera fase normalizo
se encontro medidor quemado en bornera. se normaliza colocando medidor numero XXXXXXX. se realiza inspeccion completa y se coloca tapa capilla chica. se dejo presintado.	encontro medidor quemado bornera normalizo colocando medidor numero realizo inspeccion completa coloca tapa capilla chica dejo presintado
Se Quitó Fraude Fase X Fase. Se Normalizó El Servicio. Se Realiza Acta De Fraude.	quito fraude fase fase normalizo servicio realizo acta fraude

Figure 4: Comparación de textos originales vs corregidos.

Entrenamiento del clasificador

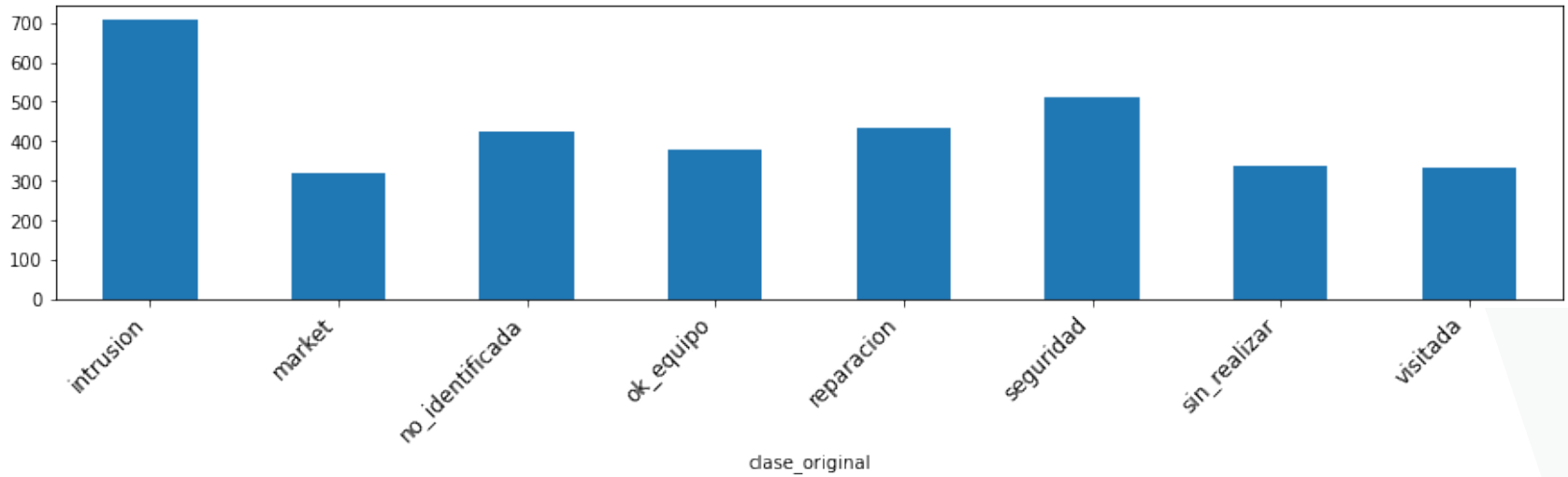
Finalmente se hizo un análisis de cantidad de palabras en los textos depurados.



Hay muy pocos textos con más de 25 palabras, por lo que se utilizó una **cota superior** para la **cantidad de palabras** a modo de **filtro**. Esto es **solamente** para el dataset de **entrenamiento**. Se puede pensar como un **hiperparámetro**.

Entrenamiento del clasificador

Como dijimos anteriormente las clases **no estan balanceadas**.



La mayoria de los samples son del tipo **Intrusion**. Las clases más relevantes son **Intrusión**, **Seguridad** y **Reparación**.

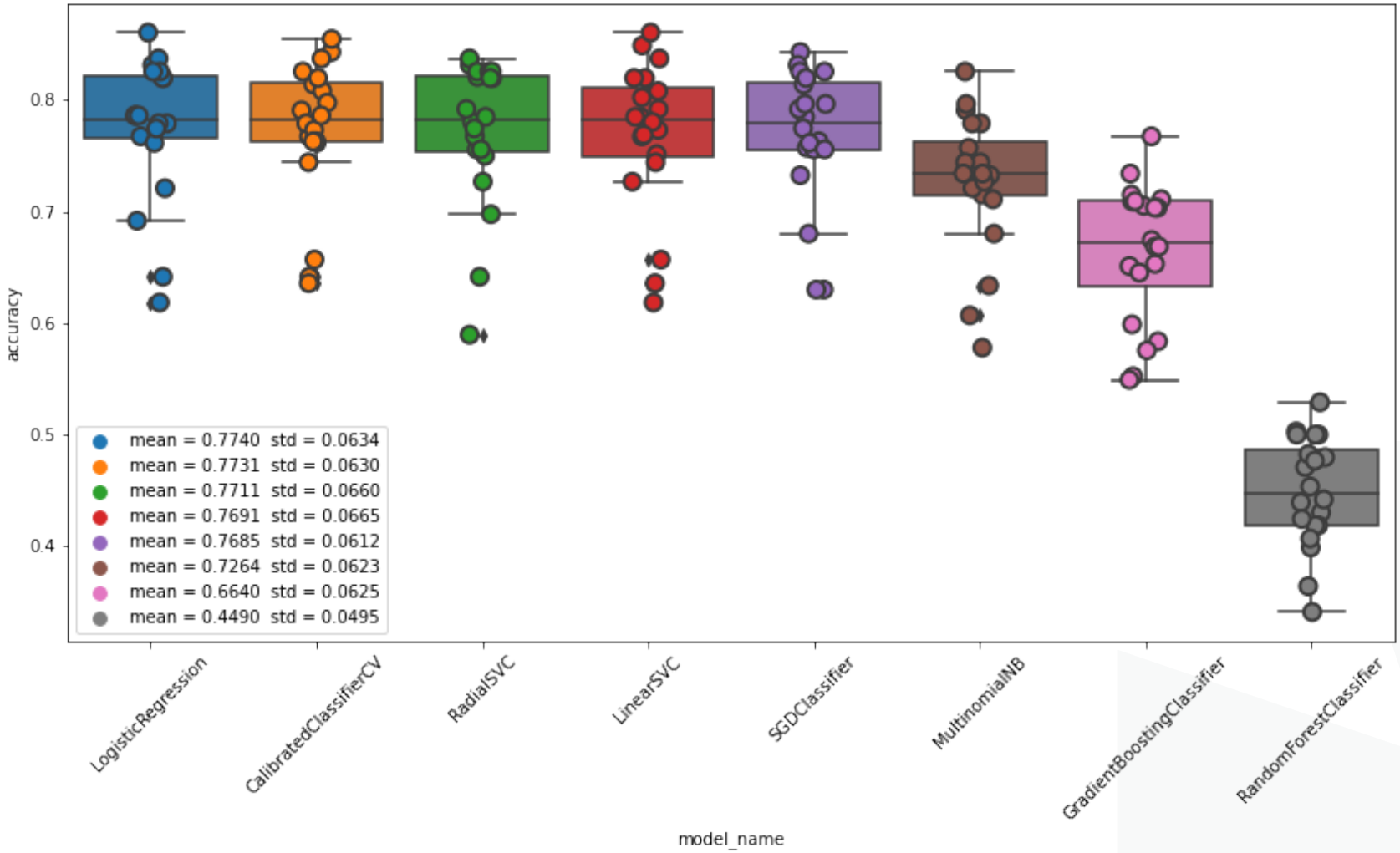
Entrenamiento del clasificador

Algoritmos:

- Regresion Logistica.
- Regresion Logistica Calibrada.
- SVC Radial.
- SVC Lineal.
- Modelo Lineal SGD
- Naive Bayes Multinomial.
- Gradient Boosting.
- Random Forest.

Entrenamiento del clasificador

Graficos de Rendimiento: Este grafico compara el rendimiento del promedio de 20 instancias de cada tipo de clasificador.

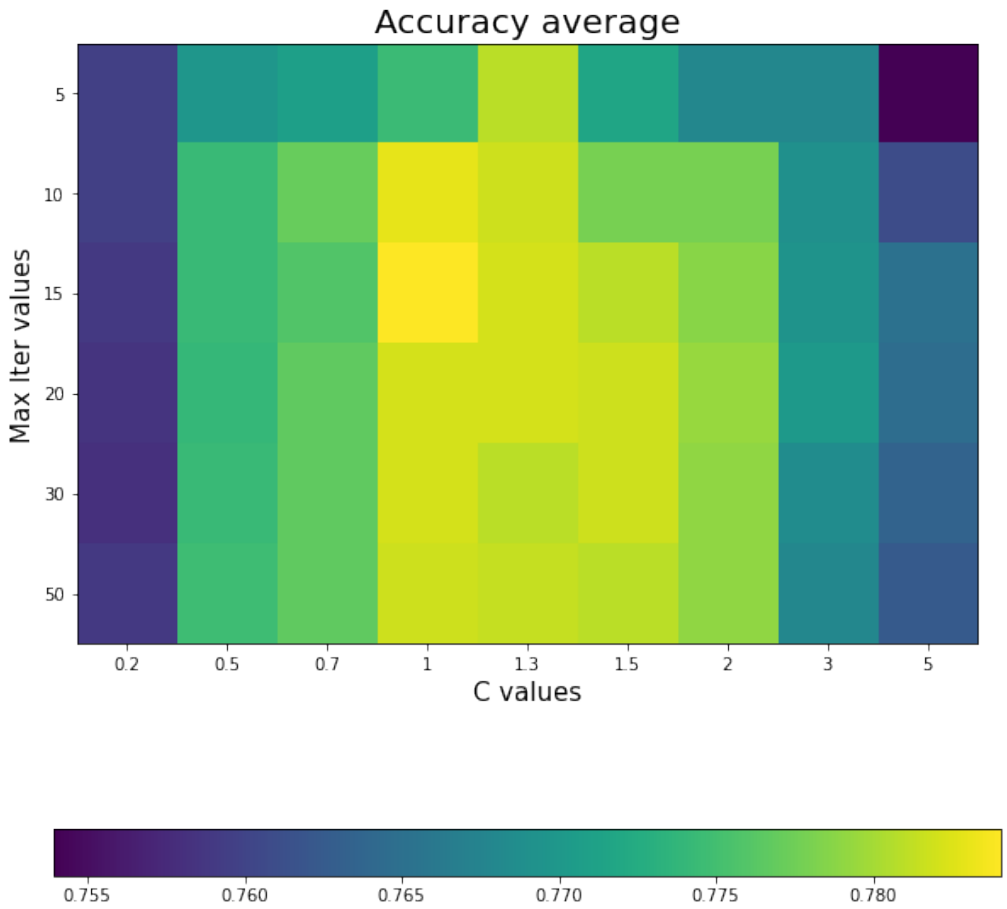


Los clasificadores de **Regresion logistica/Ccalibrada** son ligeramente mejor, seguido por los modelos **SVC radial/Lineal**.

Entrenamiento del clasificador

Grid Search para SVC Lineal:

Haciendo Grid Search sobre los parametros mas influyentes:

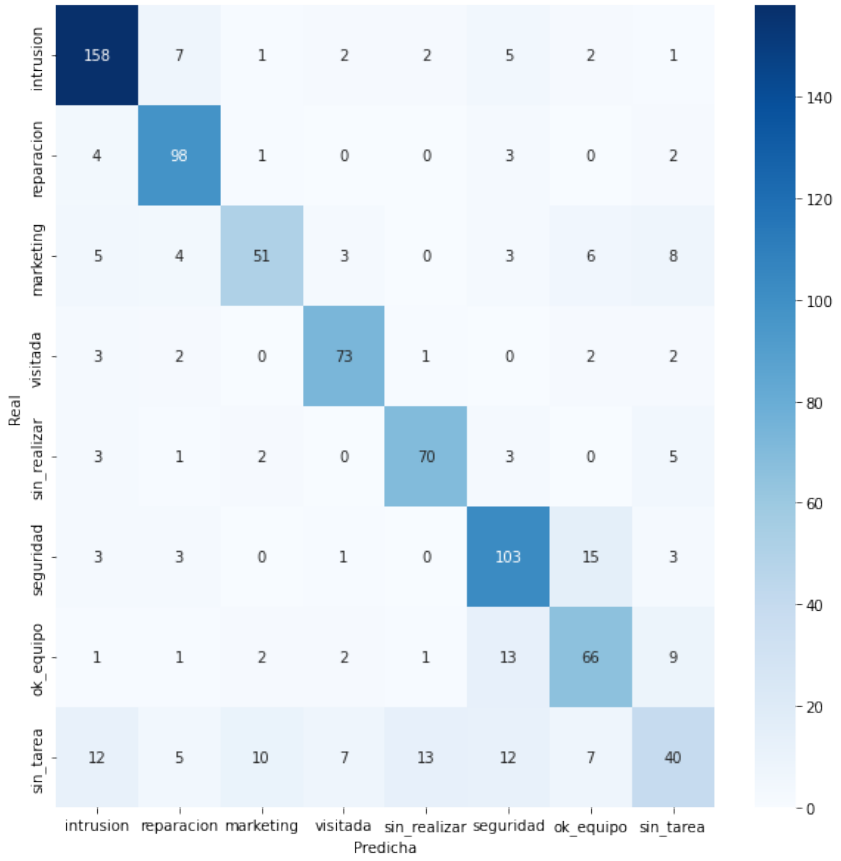


Los mejores parametros son $\{C : 1, loss : hinge, max\_iter : 15\}$.

Resultados

Matriz de confusion:

Con los mejores parametros entrenamos un clasificador **SVC Lineal** y luego analizamos los resultados en el **conjunto de prueba**.



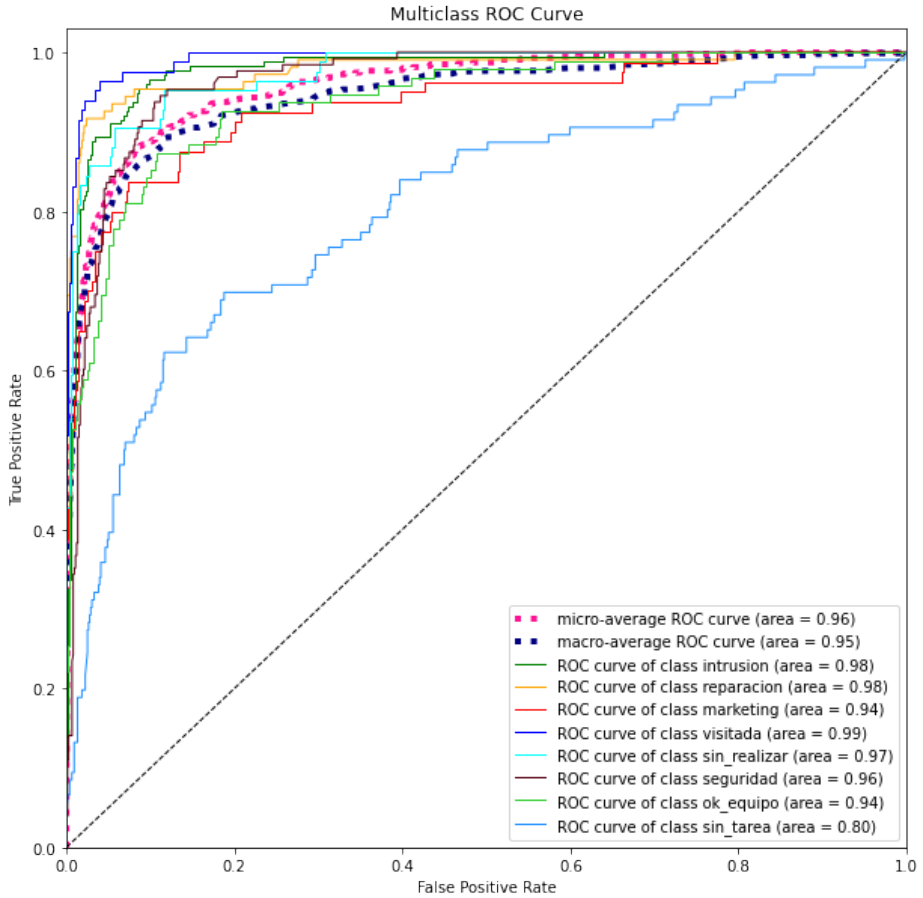
Podemos ver que el clasificador clasifica bien a casi todas las clases a excepcion de **Sin tarea** lo cual tiene sentido por ser una clase **"comodín"**.

Metricas:

Este grafico de metricas nos dice la **Precision, Sensibilidad** y **F-value** por cada clase.

Curvas AUC ROC:

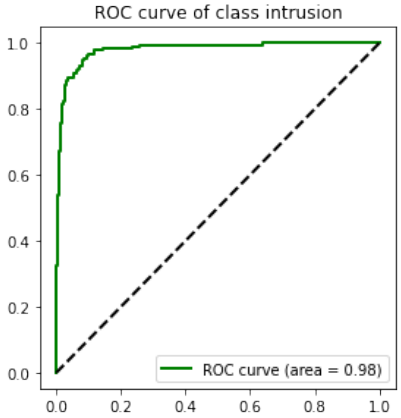
Veamos las curvas Roc para cada clase y su promedio micro/macro.

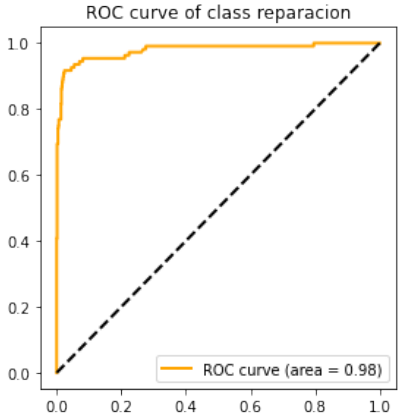


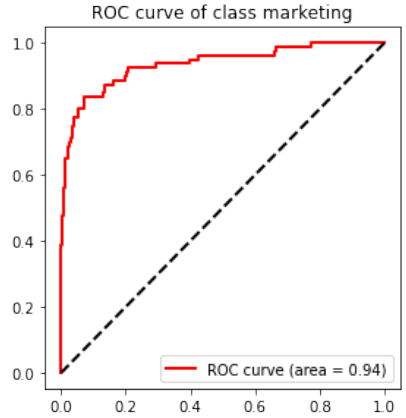
Como era de esperar la curva para la clase **Sin tarea** es considerablemente peor que el resto.

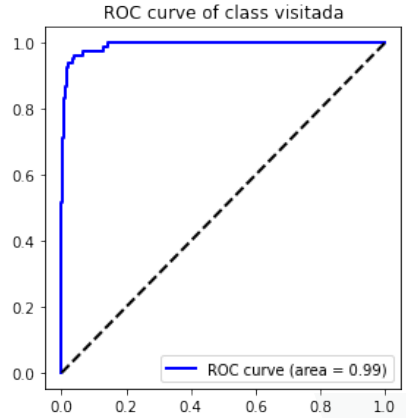
Curvas AUC ROC:

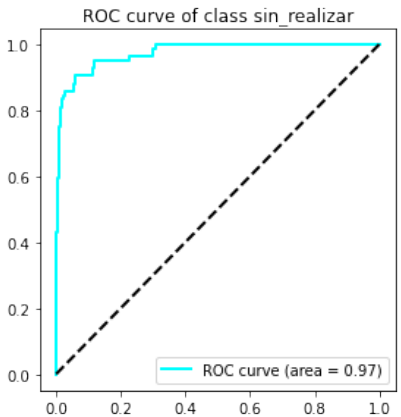
Curvas roc de cada clase (ovR).

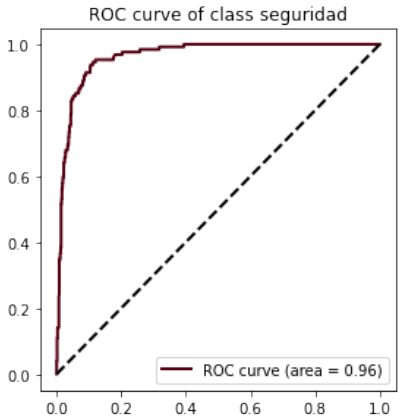


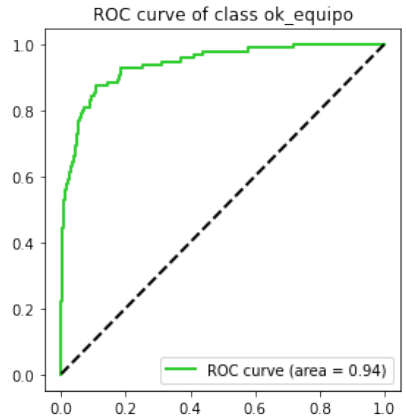


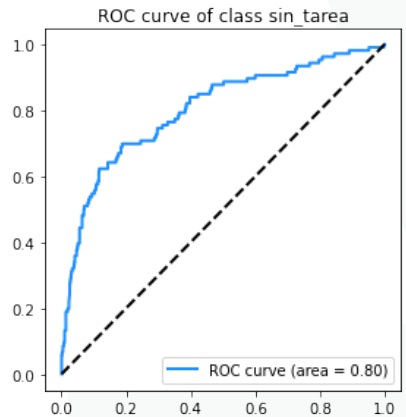












Resultados

Predicciones:

	clase_original	texto_original	texto_etl	texto_corregido	id_clase	id_predicho
2559	seguridad	medidor normal. se cambio pipeta y se coloco caja anti hurto.	medidor normal cambio pipeta coloco caja anti hurto	medidor normal se cambio pipeta y se coloco caja anti hurto	5	5
308	market	se normalizo rh directo en habitaculo cliente no posee el documento	normalizo rehabilitado directo habitaculo cliente no tiene documento	se normalizo rehabilitado directo en habitaculo cliente no posee el documento	2	0
52	intrusion	se verifico medidor monofasico sin precinto de carcasa y con el disco trabado,se procedio al cambio de medidor, tapa y termica	verifico medidor monofasico sin precinto carcasa con disco trabado procedio cambio medidor tapa termica	se verifico medidor monofasico sin precinto de carcasa y con el disco trabado se procedio al cambio de medidor tapa y termica	0	0
398	intrusion	se corto directo de linea. se normalizo instalacion de cliente. se dejo medidor en uso, se verifico lindero y se cambio tapa	corto directo linea normalizo instalacion cliente dejo medidor uso verifico lindero cambio tapa	se corto directo de linea se normalizo instalacion de cliente se dejo medidor en uso se verifico lindero y se cambio tapa	0	0
3661	visitada	ya realizada se tomo datos de medidor ya cambiado fotos en adjunto .	ya realizada tomo datos medidor ya cambiado foto adjunto	ya realizada se tomo datos de medidor ya cambiado fotos en adjunto	3	3
2259	seguridad	realizo inspeccion y control in situ con un resultado de X,XX% de error. si coloco tapa y termica, no se encontro anomalias	realizo inspeccion control interno sitio con resultado error coloco tapa termica no encontro anomalia	realizo inspeccion y control interno sitio con un resultado de de error si coloco tapa y termica no se encontro anomalias	5	5
2591	ok_equipo	(obs. form. acciones: se inspeccino med sin anomalia)	observacion inspeccion medidor sin anomalia	observacion form acciones se inspeccion medidor sin anomalia	6	6

Figure 5: Predicciones de nuevas muestras **"Id_Clase"** vs **"Id_predicho"**.

Resultados

Futuros Pasos:

- Transformers `https://huggingface.co/`
- BERT Bidirectional Encoder Representations from Transformers
- GPT-3 `https://openai.com/blog/openai-api/`
- Speech Recognition “speech-to-text”

¡Gracias por escuchar!