Pestaña 1

#### 

#### **PROYECTO FINAL DE MACHINE LEARNING**

#### 

#### **Emails de Enron con tres casos de uso**

Constanza Segrelles Munárriz  
Daniel Bravo Garcia  
Pablo Benavides Benavides

**ÍNDICE**

[Introducción 3](#_a35t66umuycc)

[Análisis de correos de Enron para la predicción de burnout laboral 3](#_8uoqryqzuysi)

[Motivación 3](#_z6dbowb9k8l3)

[1. Carga y limpieza del dataset 4](#_njzz65dtvb5l)

[2. Parsing de cada email 4](#_d8kmpd5cncxj)

[3. Limpieza de texto 4](#_hypjz55at4mt)

[4. Análisis temporal de los patrones de envío de correos 5](#_4lg17jxut1vp)

[5. Análisis de relaciones 6](#_g9t2d5vnewrh)

[6. Comparación Horario Laboral vs. Fuera de Horario 6](#_xvao70i1xh9k)

[7. Análisis de desapariciones que pueden implicar burnout 7](#_osnp7hah447b)

[8. Análisis emocional de emails con modelos de Hugging Face 8](#_jmodzarp3uih)

[9. Análisis de los desaparecidos 9](#_fqprxeauh52r)

[10. Etiquetado 9](#_ei6kaj4hp1s5)

[11. Predicción con tres modelos 10](#_ty54kvme28bq)

[Discusión y Conclusión 11](#_8zdkcl8tfus0)

[Segmentación de departamentos por clusters y su lenguaje](#_4uzvoeley77z)…………………………………..12

[1. Limpieza de datos. 12](#_vad37i3l98lw)  
 2. Creación del grafo 12

[3. Análisis de Liderazgo por Departamento (Cluster de Comunicación) 13](#_2575w639ybdf)  
 4. Comportamiento de comunicación por departamento 13

[Inferencia de urgencia e importancia de los correos electrónicos de Enron 17](#_wqjomxxuty4o)

1. [Carga y limpieza del dataset 17](#_ntw0bz4jqe6j)
2. [Enriquecimiento temporal y estructural. 17](#_2hch8kkzbzj6)
3. [Etiquetado heurístico inicial 18](#_47bgaj2mq2y5)
4. [Extracción de embeddings semánticos con DistilBERT 18](#_r6f9k2sjbrxx)
5. [Preparación del conjunto de entrenamiento y holdout, y normalización y combinación de features 18](#_al7vfj9g23il)
6. [Entrenamiento de modelos 18](#_ax1mkq5m4puy)
7. [Validación cruzada y evaluación clásica. Análisis de errores críticos y léxico 18](#_fsl2amd8l7ek)
8. [Evaluación Final del Modelo sobre el Conjunto de Validación Externa (Holdout) 18](#_sh2tfy4ij6us)

[Anexo 20](#_ca1ga71sa8qd)

# 

# 

# 

## **Introducción**

Enron fue una gran empresa energética estadounidense y, durante su quiebra a inicios de los 2000, sus correos internos fueron publicados por la Comisión Federal de Regulación Energética. Este dataset contiene ~150 usuarios (sobre todo directivos) y ~0.5 millones de correos reales. Es una de las pocas colecciones públicas sustanciales de “emails reales” corporativos [cs.cmu.edu](https://www.cs.cmu.edu/~enron/#:~:text=I%20am%20distributing%20this%20dataset,actions%20which%20precipitated%20the%20investigation). Analizar este archivo puede aportar información útil. En particular, en este informe se exploran tres ramas de análisis y aplicación: la predicción de burnout laboral, la segmentación de departamentos y la urgencia inferida. En cada sección se explicará de forma independiente las estrategias y técnicas empleadas para buscar los distintos objetivos.

# **Análisis de correos de Enron para la predicción de burnout laboral**

## **Motivación**

El “**burnout”** laboral (agotamiento laboral) es un tipo de estrés relacionado con el trabajo que provoca un desgaste físico y emocional progresivo ([mayoclinic.org](https://www.mayoclinic.org/es/healthy-lifestyle/adult-health/in-depth/burnout/art-20046642#:~:text=El%20agotamiento%20laboral%20es%20un,sentirse%20in%C3%BAtil%2C%20impotente%20y%20vac%C3%ADo)). Este síndrome ha captado gran atención por sus severas consecuencias: afecta la salud física y mental de los trabajadores y genera costes económicos sustanciales para las organizaciones y la sociedad. Por ejemplo, se estiman pérdidas superiores a los 120 mil millones de dólares anuales sólo en EE. UU. debido a problemas asociados al burnout ([journals.plos.org](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193966#:~:text=Given%20the%20negative%20effects%20of,and%20in%20multiple%20geographical%20and)). Detectar de forma temprana el “burnout” es complejo con métodos tradicionales (cuestionarios, entrevistas), por lo que se explora el uso de fuentes de datos digitales. En particular, los patrones de comunicación electrónica pueden reflejar el nivel de estrés laboral. El análisis de correos electrónicos de empleados permite identificar personas con alto riesgo de burnout tiene una potente aplicación de negocio si consigue predecir para poder prevenir este creciente malestar.

El conjunto de datos de correos de Enron es un caso de estudio ideal para investigar esta idea. La elevada carga de trabajo y tensión en Enron pueden haber generado patrones detectables en los correos enviados por sus empleados.

La siguiente sección describe en detalle el análisis realizado sobre el conjunto de correos electrónicos internos de Enron con el fin de **identificar señales de burnout laboral** (ver notebook *ML\_burnout.ipynb*). Se debe considerar como una prueba de concepto (PoC), ya que no se tenían los recursos computacionales para trabajar con todos los datos. Se explica el procesamiento paso a paso: carga de datos, parseo de campos relevantes, limpieza y enriquecimiento de los textos, análisis exploratorio de patrones temporales y relacionales, etiquetado heurístico de casos de posible burnout, y finalmente la predicción mediante modelos de Machine Learning. Cada apartado detalla el propósito de los pasos, las transformaciones aplicadas y la interpretación de resultados.

## **1. Carga y limpieza del dataset**

En esta fase se busca transformar la estructura de carpetas del dataset de Enron —originalmente organizada por usuarios y buzones— en una colección consolidada de correos individuales. El objetivo es construir un dataframe que contiene los metadatos esenciales de cada email (remitente, destinatario, fecha, asunto, cuerpo del mensaje), lo cual permitirá realizar análisis temporales, relacionales y lingüísticos en las siguientes etapas del proyecto.

El primer paso consistió en cargar el fichero comprimido con los correos de Enron. Se montó el Google Drive en Colab, se instalaron las dependencias (librerías estándar como pandas, numpy, networkx, etc.) y se definió la ruta al archivo de Enron. Con los archivos disponibles en el sistema de archivos, se procedió a listar todos los correos. Para ello se definió la función leer\_todos\_los\_archivos(path\_destino) que recorre recursivamente la estructura de directorios (usuario → carpeta → archivo) y construye una lista de registros con las rutas completas de cada correo, junto con metadatos como el nombre de usuario, carpeta y nombre de archivo. Al ejecutar esta función se obtuvo el número total de correos encontrados (por ejemplo, 30605 correos en Enron) y se mostraron algunos ejemplos de rutas. Esta tabla inicial permite tener una idea de la magnitud de datos a procesar y sirve de entrada para las siguientes etapas. En particular, se almacena la lista en todos\_los\_archivos para su uso posterior.

## **2. Parsing de cada email**

Esta sección tiene como finalidad extraer y estructurar de manera adecuada la información contenida en cada correo. Dado que los emails vienen en texto plano con formatos heterogéneos, se requiere un procesamiento específico para identificar de forma robusta campos clave como la fecha, remitente, destinatarios y el contenido del mensaje. Esto es clave para poder vincular los emails con patrones de comportamiento temporal, emocional y social.

Con la lista de archivos, se **parseó cada correo** para extraer sus campos clave. Se definió la función parsear\_email(texto) que recibe el texto plano del correo y usa expresiones regulares para buscar cabeceras estándar (From:, To:, Cc:, Date:, Subject:). Cada coincidencia se almacena en un diccionario de campos (from, to, cc, date, subject). Luego se separa el cuerpo del mensaje (después de la primera línea en blanco doble) y se guarda en el campo body. Un bucle con tqdm itera sobre los correos (en la versión de prueba se limitado a 1000 para rapidez) abriendo cada archivo con codificación Latin1 para evitar errores, leyendo el texto y aplicando la función de parseo. Para cada correo, además de los campos extraídos, se agrega un registro con información del archivo (usuario, carpeta, nombre) para trazabilidad. El resultado es una lista correos\_parseados con un diccionario por correo. Finalmente se crea un DataFrame df de esta lista y se guarda en un CSV (correos\_parseados.csv) para agilizar futuras recargas.

## **3. Limpieza de texto**

Una vez extraído el contenido de los emails, es necesario normalizarlo para su posterior análisis lingüístico y de sentimiento. El objetivo aquí es eliminar ruido, como encabezados repetitivos, firmas automáticas o caracteres extraños, y aplicar técnicas de procesamiento de lenguaje natural como lematización y eliminación de stopwords. Esto permitirá capturar de manera más precisa el tono y el contenido emocional de los correos.

El siguiente objetivo fue **preparar los datos para análisis de NLP**. En primer lugar, se cargó el DataFrame desde el CSV y se inspeccionaron sus primeras filas, confirmando que contiene las columnas esperadas (from, to, date, subject, body, además de user, etc.). Luego se realizó la limpieza textual del cuerpo del email. Se definió la función limpiar\_texto(text) que, mediante expresiones regulares, elimina contenido de mensajes reenviados (“forwarded message”), quita URLs (http…), unifica espacios múltiples en uno solo y elimina caracteres especiales no alfanuméricos innecesarios. Aplicando esta función a la columna body se generó una nueva columna clean\_body con el texto limpio de cada email. Se verificó visualmente que las cinco primeras filas mostraran textos sin enlaces ni caracteres extraños.

Además, se convirtieron las fechas a formato datetime de Pandas: la columna original date (texto) se transformó a parsed\_date con pd.to\_datetime. A partir de esta fecha parseada se extrajeron nuevas variables temporales: año, mes, día, nombre del día de la semana (weekday) y hora. Con la hora se definió un campo categórico horario indicando si el email fue enviado en horario laboral (8–18) o fuera de él (“laboral” vs “no\_laboral”). Estas transformaciones enriquecen cada registro de correo con atributos de tiempo útiles para el análisis posterior. El resultado es un DataFrame con ~30605 filas y columnas que incluyen el cuerpo limpio, fecha/hora parseada y variables derivadas (ver salida de df.info() en secciones posteriores).

## **4. Análisis temporal de los patrones de envío de correos**

Esta sección explora el comportamiento temporal agregado del envío de emails a lo largo del tiempo. El objetivo es identificar posibles ciclos, tendencias y anomalías que puedan estar relacionadas con eventos organizacionales o estados de ánimo colectivos. Además, el análisis temporal es una herramienta crucial para detectar desviaciones que podrían ser indicativas de estrés laboral o inicios de burnout.

Con los datos temporales disponibles, se examinó **cómo varía la actividad de envío a lo largo del tiempo** (Anexo: Figura 1) y en relación con fines de semana o festivos. Se agregó la columna booleana weekend (True si el día es sábado o domingo) usando la lista de nombres de día. Además se usó la librería holidays de Python para identificar días festivos: se creó un objeto de festivos de EE. UU. y otro específico de Texas (estado sede de Enron). Luego se marcó is\_holiday si la fecha del email coincide con alguno de estos días festivos de Texas. Con estas banderas, se definió un indicador off\_hours igual a True cuando la hora es antes de 9 o después de 18 (fuera de jornada), y non\_working\_day si es fin de semana o festivo.

A continuación, se agruparon los correos por usuario y fecha para contar cuántos enviaban fuera de horario cada día (daily\_user\_activity). También se hizo una agregación general diaria total de correos fuera de horario (daily\_off\_hour\_counts). Como ejemplo de visualización, se trazó un gráfico de líneas de la serie temporal de la cantidad diaria de correos fuera de horario de los 30 usuarios con masyor número de emails enviados (Anexo: Figura 2). .

## **5. Análisis de relaciones**

Aquí se pretende hacer un superficial mapeo de las relaciones interpersonales dentro de la organización utilizando la información de los correos enviados y recibidos. Se construyen grafos que permiten observar cómo se estructuran las redes de comunicación, quiénes son los nodos centrales y cómo fluye la información. Estos patrones relacionales pueden ofrecer indicios sobre niveles de carga laboral, aislamiento o sobreexposición de ciertos individuos.

Se construyó un **grafo de comunicación** entre empleados para identificar patrones de interacción. Usando networkx, se creó un grafo dirigido G. Se recorrió cada correo del DataFrame completo, y para cada remitente (from) y sus destinatarios (to se dividió por comas), se agregó una arista en el grafo. Cada vez que un mismo par remitente-destinatario aparece, se incrementa el peso (número de correos enviados). Tras procesar todos los correos, se imprimieron estadísticas básicas de la red: por ejemplo, *6646* nodos (usuarios únicos) y *12839* aristas (envíos de correo) en la red total. Esto indica la complejidad de la red de comunicación. Luego se calculó la **centralidad de grado** de cada nodo (medida de cuántos enlaces salen o entran a un usuario) y se ordenaron los 10 usuarios más centrales. Entre ellos destacaron nombres como *jeff.dasovich@enron.com* (con centralidad ~0.1505), *michelle.cash@enron.com* (0.0891), *rick.buy@enron.com* (0.0670), etc. Esto revela quiénes son los empleados más conectados en el flujo de correos.

Para visualizar globalmente la red se intentó dibujar el grafo con nx.draw(), aunque con miles de nodos la imagen resultante es muy densa y de poco detalle (solo sirve para apreciar la densidad general - Anexo: Figura 3). En resumen, este análisis de relaciones permite identificar estructuras organizacionales o aislados: por ejemplo, si un nodo importante deja de enviar correos podría ser señal de que está desconectado del flujo informativo, lo cual podría relacionarse con **síntomas de burnout** (aislamiento o baja participación en la red).

En la red analizada (Anexo: Figura 4 y Tabla 1), algunos usuarios presentan un alto *in-degree* pero *out-degree* cero, como *richard.shapiro@enron.com* o *william.kendrick@enron.com*. Esto significa que reciben correos de varios remitentes (aproximadamente el 4.1% de los nodos), pero no envían mensajes, lo que sugiere un rol pasivo, posiblemente como supervisores o receptores de reportes. Por otro lado, usuarios con alto *out-degree* y sin *in-degree*, como *rick.buy@enron.com*, envían muchos correos sin recibir respuesta. Esto indica un rol muy activo fuera del horario laboral, probablemente liderando, delegando tareas o comunicando instrucciones, lo que podría asociarse con una mayor carga de trabajo y potencial riesgo de burnout. Además, en este subconjunto, el betweenness centrality es cero para todos los nodos, lo que implica que no hay intermediarios en la red: las comunicaciones fluyen directamente de emisor a receptor. Esto refleja una estructura jerárquica y distribuida, sin nodos que centralicen o canalicen la información.

## **6. Comparación Horario Laboral vs. Fuera de Horario**

El objetivo de esta sección es analizar cómo varía el envío de correos en función del horario laboral. La hipótesis es que una frecuencia alta de correos enviados fuera del horario habitual puede ser un signo de sobrecarga o mala gestión del tiempo. Se comparan patrones diarios y semanales, buscando diferencias significativas que puedan estar asociadas con estrés o agotamiento.

Para cuantificar qué empleados envían más correos fuera de horario, se realizó un análisis tabular y gráfico. Primero se agruparon los datos por usuario y la variable booleana off\_hours (True = fuera de horario, False = dentro de horario). Con groupby(['user','off\_hours']).size().unstack() se obtuvo una tabla de recuentos por usuario, rotulada como “Laboral” (False) y “Fuera\_de\_horario” (True). Se agregaron columnas de total de correos y porcentaje de correos fuera de horario por usuario. A partir de allí se seleccionaron los 15 usuarios con mayor porcentaje de emails fuera de horario (top\_15\_out). Este ranking mostró, por ejemplo, que *guzman-m* envió 3 correos y todos (100%) fuera del horario laboral; otros casos fueron *crandell-s* con ~86% fuera; *corman-s* con ~47%, *cuilla-m* con ~46%, *buy-r* ~43%, etc. (ver Anexo: Tabla 2).

Como visualización, se creó un gráfico de barras apiladas (ver Anexo: Figura 5) donde en el eje horizontal aparecen los 15 usuarios top, y en cada barra se comparan las cantidades de correos “Laboral” vs “Fuera\_de\_horario”. Esto facilita ver, no solo el porcentaje, sino el volumen absoluto. La figura resultante destaca que, por ejemplo, *cash-m* envió un alto volumen de correos en general (~1880 laborales y 1045 fuera), mientras que *crandell-s* envió muchos correos fuera de horario respecto al total.

Estos resultados permiten identificar casos extremos: un empleado que mantiene el 100% de su comunicación fuera de horario o alguien con un alto volumen nocturno. Tales patrones pueden indicar que compensan carga de trabajo excesiva con actividad extra (trabajan a deshoras), lo cual es un posible síntoma de estrés laboral elevado o burnout.

## **7. Análisis de desapariciones que pueden implicar burnout**

Esta etapa desciende al nivel de análisis individual para observar cómo varía la actividad de cada empleado a lo largo del tiempo. Se busca identificar patrones atípicos, caídas abruptas en la actividad o cambios de comportamiento que puedan reflejar una situación de burnout incipiente. Este análisis personalizado permite construir perfiles dinámicos de comportamiento laboral.

A continuación, se analizó en detalle la actividad diaria de cada usuario seleccionado. Se filtraron aquellos usuarios que tenían más del 20% de sus correos fuera de horario, obteniendo una lista de usuarios a monitorear (por ejemplo: *brawner-s, buy-r, campbell-l, carson-m, cash-m, causholli-m, corman-s, crandell-s, cuilla-m, dasovich-j, guzman-m*). Para esos usuarios se extrajeron únicamente los correos fuera de horario y se contaron cuántos correos envió cada día.

Luego, para cada usuario se trazó la serie temporal de correos diarios fuera de horario. En cada subgráfico (un usuario por fila) se muestran los puntos de actividad diaria. Además, para detectar picos anómalos, se trazó una línea de umbral: media + 2·desviación estándar de su serie de conteos. Esto evidencia días en los que un usuario disparó un número inusualmente alto de emails (los picos por encima del umbral punteado). Este *ejemplo de detección de picos anómalos* ayuda a visualizar momentos críticos: si un empleado se comunica mucho en un día (pico) y luego deja de hacerlo bruscamente, podría reflejar momentos de sobrecarga seguidos de retirada. Dos ejemplos del gráfico múltiple (Anexo: Figura 6) ilustra estas trayectorias diarias para cada usuario seleccionado. Se observó, por ejemplo, que *carson-m* tuvo ciertos días con más de 10 correos fuera de horario (muy por encima de su media de ~3.5), luego de los cuales su actividad diurna se redujo drásticamente. Este patrón de “pico-anomalía” podría interpretarse como que el empleado trabajó intensamente en un momento y luego entró en reposo, señalando estrés acumulado. Por tanto, la detección automática de estos picos se planteó como indicador: días con actividad inusual alta podrían preceder a una posible “desaparición” de correos, que entra en el siguiente análisis.

Ahora se exploran los períodos de inactividad **prolongada** por parte de empleados previamente activos. La hipótesis es que una desaparición comunicacional abrupta podría indicar una baja por enfermedad, una renuncia o incluso un caso de burnout severo. Detectar estas ausencias puede ayudar a etiquetar comportamientos críticos retrospectivamente y alimentar modelos predictivos.

Para formalizar la idea anterior, se implementó un algoritmo de **detección de desapariciones prolongadas**. La idea es identificar usuarios que, tras un periodo de actividad alta, dejan de enviar correos por un lapso largo (lo que puede indicar burnout o ausencia). Se definieron parámetros: un umbral de desaparición de 14 días, y una ventana de “alta actividad” previa de 7 días. Solo se consideraron los correos fuera de horario. Para cada usuario, se calculó la fecha del último correo y cuántos días han pasado desde entonces hasta la fecha final del dataset. También se calculó la media y desviación de sus conteos diarios previos, definiendo “alta actividad” como un día con conteo > media + 1·desv.

Si un usuario cumple dos condiciones (*gap\_dias* ≥ 14 y tuvo al menos un día de alta actividad en los últimos 7 días antes de desaparecer), entonces se lo añadió a la lista de “posible desaparición”. Este criterio busca hallar empleados que finalizaron abruptamente con un pico de trabajo. Al ejecutar esto se obtuvo un pequeño conjunto de casos: por ejemplo *causholli-m, cuilla-m, buy-r, carson-m, crandell-s* fueron etiquetados con posible desaparición. Para cada uno se registró la fecha de último correo, días sin actividad y sus estadísticas medias. Como salida destaca que, por ejemplo, *causholli-m* desapareció en 2001-11-28 tras 6971 días de actividad (lo cual en este dataset extenso equivale a su último registro), con media diaria ~4.58 correos fuera de horario y pico antes. Todos estos usuarios resultaron con had\_high\_activity\_before\_gap=True, validando que terminaron con un día pico.

La conclusión de este análisis es que las desapariciones súbitas en el envío de correo, tras un reciente episodio intenso, pueden ser un síntoma de burnout laboral: el empleado habría experimentado agotamiento y cesó su actividad. Se grafica opcionalmente la evolución diaria de cada uno (Anexo: Figura 6), con la última fecha en rojo y el umbral de alta actividad en naranja, confirmando visualmente el patrón (se ve la caída brusca al final). Este heurístico de detección de desapariciones complementa la intuición de los picos anómalos.

## **8. Análisis emocional de emails con modelos de Hugging Face**

El objetivo de esta sección es evaluar el contenido emocional de los correos electrónicos utilizando modelos de lenguaje preentrenados. El análisis emocional permite estimar el tono de la comunicación (positivo, negativo, neutral, etc.), lo cual es fundamental para detectar estados emocionales deteriorados o signos de tensión prolongada que puedan correlacionarse con burnout.

Para explorar el contenido emocional de los correos, se empleó un modelo preentrenado de **Transformers** de Hugging Face. En particular, se usó el pipeline de clasificación de texto con el modelo j-hartmann/emotion-english-distilroberta-base, que etiqueta textos con siete categorías emocionales (anger, disgust, fear, joy, neutral, sadness, surprise). Se cargó este *pipeline* de la librería transformers y se aplicó a un subconjunto de correos de usuarios seleccionados (por rapidez, en la versión se analizó solo los mensajes de los usuarios considerados críticos del análisis anterior). Al aplicarlo, el modelo devuelve para cada correo un conjunto de puntajes de probabilidad para cada emoción.

De cada resultado se extrajo la **emoción dominante** (la etiqueta con mayor probabilidad) y se guardó en la columna main\_emotion. También se mantuvieron todos los puntajes en emotion\_scores para análisis posterior. A modo de verificación, se agrupó este subconjunto por usuario, por off\_hours (dentro/fuera de horario) y por main\_emotion, construyendo una tabla de contaje de emociones (Anexo: Tabla 3). Por ejemplo, para el usuario *buy-r* se obtuvo que en horario laboral predominan correos neutros (1068) y alegres (96), mientras que fuera de horario su distribución cambia levemente (839 neutros, 53 alegres, etc.). Esto muestra cómo varían las emociones según contexto temporal. La Figura 7 muestra un ejemplo de la evolución temporal de las emociones para un usuario.

En resumen, en esta etapa se tradujo texto de correos a un vector emocional, lo que permite cuantificar si usuarios bajo estudio manifiestan mayor carga de negatividad (p.ej. tristeza, enfado) en sus mensajes, especialmente fuera de horario. Esto alimentará el etiquetado de burnout basado en emociones negativas.

## **9. Análisis de los desaparecidos**

Esta parte busca caracterizar a los individuos identificados previamente como “desaparecidos”, analizando su comportamiento previo y comparándolo con el del resto de los empleados. El propósito es validar si existen señales consistentes que preceden a su inactividad, como mayor envío fuera de horario o contenido emocional negativo, lo cual fortalecería las hipótesis del modelo de burnout.

Los empleados identificados como *desaparecidos* en el punto anterior (*causholli-m, cuilla-m, buy-r, carson-m, crandell-s*) fueron objeto de examen adicional. Esto sugiere que estos empleados críticos merecen atención reforzada: la combinación de actividad extrema seguida de cese completo, junto con la posible prevalencia de tonos negativos en sus correos (por ejemplo, en los recuentos fuera de horario de *cuilla-m* se observó una proporción significativa de “sadness”, “anger” y “fear”), refuerza la hipótesis de burnout. El análisis muestra coherencia: los sujetos con desapariciones suelen haber tenido comunicaciones intensas y potencialmente más negativas, apoyando la idea de que sus ausencias no son aleatorias sino signos de agotamiento laboral.

## **10. Etiquetado**

### **10.1 Etiquetado por inactividad**

Se propone aquí una estrategia de etiquetado binario basada en la actividad comunicacional. Aquellos individuos con una caída sostenida en la frecuencia de correos son marcados como candidatos a burnout. Este etiquetado es esencial para poder entrenar modelos supervisados más adelante, y se basa en reglas definidas a partir del análisis temporal previo.

Como primer método heurístico de etiquetado de burnout, se definió **“burnout\_inactivity”**: un empleado se etiqueta como “burnout por inactividad” si su último correo registrado precede a una fecha límite establecida. Concretamente, se calculó para cada usuario la fecha de su último correo (last\_active\_date). Se fijó un umbral arbitrario que consideramos como señal de que dejó de comunicarse antes de esa fecha. Sobre esa base, se creó una columna booleana burnout\_inactivity que es True si last\_active\_date < threshold. Finalmente, esta etiqueta se unió al DataFrame principal por usuario.

Este enfoque “supervisado” asume que quienes dejaron de enviar correos se consideran en burnout por inactividad. Aunque simplista, permite capturar casos obvios: el DataFrame resultante (ver salida df\_subset.info()) muestra que la columna burnout\_inactivity tiene True para aquellos usuarios inactivos. Este será un label que luego se combina con otros para entrenar modelos predictivos.

### **10.2 Etiquetado por emociones negativas y correos fuera de horario**

Esta sección introduce un segundo criterio de etiquetado, esta vez centrado en indicadores indirectos de agotamiento: el uso frecuente de correos con carga emocional negativa y el envío habitual fuera del horario laboral. El objetivo es capturar comportamientos que, aunque no terminen en inactividad, representen riesgo de burnout.

Se definieron las emociones negativas de interés: *anger*, *fear*, *sadness* y *disgust*. Para cada correo ya clasificado por el modelo anterior, se calculó un “puntaje negativo” (negative\_score) sumando las probabilidades asignadas a esas emociones negativas. Luego se agruparon los datos por usuario y semana (frecuencia semanal) para obtener medidas de tendencia: en el DataFrame weekly\_agg se calculó la media semanal de negative\_score, la media de off\_hours (proporción de correos fuera de horario esa semana) y el recuento de correos (clean\_body.count).

Con estas métricas semanales se busca detectar empleados con **persistencia de negatividad** y trabajo fuera de horario. Por ejemplo, un alto avg\_negative\_score acompañado de un valor alto de avg\_off\_hours en varias semanas podría indicar estrés sostenido. Estas características constituyen indicadores temporales heurísticos: no se asigna aún una etiqueta final, pero facilitan separar usuarios con patrón de alta carga negativa.

### **10.3 Combinación de ambas etiquetas para una etiqueta más general**

Aquí se construye una etiqueta compuesta que combina las señales de inactividad, emociones negativas y trabajo fuera de horario. Este enfoque busca una representación más rica y realista del fenómeno del burnout, permitiendo capturar distintos perfiles de riesgo y mejorar la robustez del modelo predictivo que se entrenará a continuación.

Se creó entonces una etiqueta compuesta burnout\_label que combine los dos criterios previos (inactividad *o* estrés emocional). El resultado es una etiqueta final binaria por correo (y por usuario) que considera al empleado con burnout si apareció inactivo o con altas emociones negativas. Esta etiqueta general servirá como la variable objetivo en el modelado.

## **11. Predicción con tres modelos**

La sección final se enfoca en entrenar y comparar varios modelos de aprendizaje automático para predecir burnout a partir de las etiquetas generadas. Se evalúan modelos como Random Forest, XGBoost y Regresión Logística, con métricas de rendimiento como *Precision*, *Recall* y *F1-scores*. El objetivo es identificar qué modelo ofrece mejores capacidades para detectar de manera temprana y fiable situaciones de burnout entre los empleados de Enron.

Para ello, se prepararon datos para entrenar modelos de clasificación que predigan la etiqueta de burnout. Lo ideal sería predecir si un empleado está en riesgo de burnout, ya que se trata de un fenómeno acumulativo que se manifiesta en el comportamiento general del individuo y no necesariamente en un solo email. En ese caso haríamos una agrupación por user, ya que la unidad de análisis es la persona y no los emails individuales. Sin embargo, debido a la falta de capacidad computacional (no somos capaces de acceder a todos los usuarios con sus múltiples emails) y debido también a que seguramente 150 usuarios no serían suficientes para hacer un modelo predictivo, nos enfocamos en hacer una predicción de señales de burnout a partir de cada email usando solo una muestra. En este caso, no es necesario agrupar: se puede construir un modelo a nivel de email, utilizando cada correo como una muestra individual. Se separaron los features (X) y el target (y = burnout\_label). Se dividió en conjunto de entrenamiento y prueba (20%) con train\_test\_split. Se definieron tres modelos: Random Forest, Regresión Logística y XGBoost, se entrenaron y se evaluaron. Además, al tener los datos del target poco balanceados, burnout (clase minoritaria) vs no burnout (clase mayoritaria), se aplicó la técnica SMOTE en los datos de entrenamiento que genera ejemplos sintéticos de la clase minoritaria para que el modelo tenga suficiente información y no la "ignore".

Los resultados (Anexo: Tabla 4) (6023 muestras) muestran un rendimiento sólido, especialmente en los modelos Random Forest y XGBoost, que alcanzan una **accuracy general del 93–94%** y **f1-scores elevados en la clase mayoritaria (no burnout)**. Esto refleja una capacidad robusta para capturar patrones generales del conjunto de datos. Sin embargo, el desempeño en la clase minoritaria (burnout) sigue siendo modesto: el **Random Forest logra un f1-score de 0.28** y XGBoost **de 0.24**, con recalls de apenas **0.38 y 0.37**, respectivamente. Esto indica que, aunque algunos casos de burnout son detectados, el modelo aún omite una proporción considerable. Esta limitación se debe en gran parte al **desbalance natural del dataset** (ya que SMOTE no se aplicó a el conjunto de test), donde los casos “burnout” representan menos del 4%. Aun así, este desbalance es representativo de escenarios reales en los que el burnout es poco frecuente pero crítico de identificar. En conjunto, los modelos son funcionales como primera aproximación, pero aún hay margen de mejora para elevar la sensibilidad sin deteriorar gravemente la precisión.

## **Discusión y Conclusión**

Los resultados deben interpretarse con precaución. El etiquetado de “burnout” es heurístico, basado en patrones como ausencia de correos o emociones negativas, no en diagnósticos clínicos reales. Esto implica que los indicadores usados son aproximaciones, no pruebas definitivas de agotamiento laboral. El análisis se basa en un subconjunto limitado de correos internos de Enron, con posibles sesgos de selección (directivos, datos anonimizados o incompletos). Además, otros factores podrían explicar patrones detectados, como cargas laborales temporales o renuncias. El modelo de emociones utilizado tiene una precisión moderada (~80%) y puede cometer errores, aunque combinarlo con variables temporales mejora la robustez del análisis.

Pese a estas limitaciones, los hallazgos son coherentes con estudios previos: patrones como correos en horarios atípicos y lenguaje negativo se asocian con burnout. Esto sugiere que herramientas similares podrían servir como sistemas de alerta temprana en entornos reales. Sin embargo, cualquier uso práctico debe considerar implicaciones éticas, como privacidad y consentimiento de los empleados. Finalmente, el estudio abre camino a mejoras técnicas: incluir nuevas variables, usar mejores modelos NLP, o validar con datos reales (como encuestas de burnout) para aumentar la confiabilidad del modelo.

El notebook *ML\_burnout.ipynb* contiene la investigación exploratoria de este trabajo. Muestra cómo, a partir de un gran volumen de correos electrónicos, se puede construir una herramienta para detectar indicios de burnout. Se aplicaron técnicas de limpieza de texto, extracción de variables temporales, análisis emocional con NLP y etiquetado heurístico. Luego, se entrenaron modelos de clasificación (Random Forest, Regresión Logística, XGBoost) y se evaluaron sus predicciones.

Los aprendizajes clave son: (1) la combinación de volumen de correos y emociones mejora la detección de riesgo; (2) los análisis exploratorios permiten identificar cambios de comportamiento; y (3) sin validación externa, los resultados deben interpretarse con cautela.

En conjunto, el análisis demuestra el potencial de aplicar ciencia de datos a comunicaciones internas para identificar señales de agotamiento, siempre con acompañamiento ético y experto.

# **Segmentación de departamentos por clusters y su lenguaje** El conjunto de datos de los correos, contiene campos como from, to, cc, user, subject, body, así como metadatos de fecha, hora, día de la semana y número de destinatarios. En esta parte vamos a analizar la cultura lingüística empresarial.

## **1. Limpieza de datos**

Se utilizó la base de datos elaborada en el análisis de correos de Enron para predecir el burnout laboral, aplicando una limpieza previa para asegurar el correcto funcionamiento del código.

La columna *TO* fue transformada en listas homogéneas de destinatarios mediante funciones que manejan valores nulos, listas en formato cadena y arrays de NumPy.

A partir de esto, se construyó un grafo no dirigido donde los nodos representan usuarios y las aristas, ponderadas por frecuencia, indican vínculos de comunicación. El grafo resultante incluye 6,675 nodos y 12,524 aristas, mostrando una red extensa con un núcleo de usuarios de alto intercambio y múltiples conexiones esporádicas, lo que sugiere subgrupos o comunidades dentro de una estructura diversa.

## **2. Creación del grafo** Tras aplicar un algoritmo de detección de comunidades basado en modularidad, se generó un layout en dos niveles: las comunidades se ubicaron en los vértices de un gran círculo para resaltar su separación, y dentro de cada bloque se utilizó un algoritmo de resorte que distribuye las posiciones internas sin solapamientos excesivos, manteniendo la cohesión del grupo.

Esta disposición revela una estructura modular clara: comunidades bien definidas, con núcleos densos de alta actividad y perímetros difusos que enlazan distintos grupos, lo que sugiere una coexistencia de agrupaciones cerradas y conexiones intercomunitarias.

Con los datos agrupados por similitud en función del intercambio de correos, es posible comenzar el análisis de los clusters, como se muestra en la Figura 8: Grafo de departamentos por comunicación.

## **3. Análisis de Liderazgo por Departamento (Cluster de Comunicación)**

En esta primera etapa se identificaron comunidades dentro del grafo de comunicación de Enron, interpretadas como departamentos u órganos funcionales.  
 A partir de las métricas de red se extrajo el Top 3 de usuarios más relevantes por cada indicador en cada comunidad, lo que permite varias lecturas:

* **Líderes de actividad (Top Grado):** suelen ser coordinadores que concentran solicitudes, reportes o aprobaciones, reflejando su rol clave en la operación diaria.
* **Líderes de influencia (Top PageRank):** reciben comunicación de otros nodos influyentes, lo que los vincula con cargos directivos o técnicos estratégicos. Son esenciales en la difusión de mensajes clave y cambios organizativos.
* **Puentes estratégicos (Top Betweenness):** conectan grupos aislados al aparecer en rutas críticas, favoreciendo la colaboración transversal y evitando silos informativos.

Este análisis de liderazgo por comunidad no solo identifica a los usuarios más activos, influyentes y conectores, sino que también ofrece una base para optimizar la comunicación interna, detectar cuellos de botella e impulsar planes de sucesión o formación centrados en los verdaderos hubs de la red.

## **4. Comportamiento de comunicación por departamento**

En el segundo paso se analizó el estilo de comunicación de cada departamento (cluster) mediante métricas agregadas de los correos asignados: hora promedio de envío, porcentaje de correos en horario laboral, número medio de destinatarios por mensaje y volumen total de mensajes enviados por grupo.

Departamentos con muchos destinatarios por correo suelen tener funciones de difusión interna, mientras que en equipos más reducidos la comunicación tiende a ser más directa y focalizada.

Estos indicadores permiten identificar distintos estilos de trabajo: grupos con comunicación continua, departamentos de difusión masiva, equipos nocturnos o internacionales, y otros centrados en momentos específicos.

En esta fase se extrajeron las variables clave: hora promedio de envío, porcentaje de correos laborales, destinatarios promedio y total de correos enviados.

## **5. Comparación de lenguaje entre departamentos y topics.**

En el tercer paso se construyó un vocabulario común de la organización aplicando TF-IDF sobre el cuerpo completo de correos, excluyendo tokens numéricos y términos genéricos. Esto permitió identificar las 300 palabras más relevantes a nivel global (por ejemplo: *meeting*, *power*, *market*, *energy*, *report*). Luego, se calculó el TF-IDF medio por departamento y se extrajeron las 100 palabras más distintivas de cada grupo, revelando su valor lingüístico propio.

En el cuarto paso se realizó un análisis semántico avanzado para identificar los temas frecuentes en cada departamento. Se trabajó solo con comunidades con al menos 20 mensajes, asegurando representatividad estadística.  
Para hacer comparaciones consistentes, se construyó un modelo de bolsa de palabras, filtrando tanto términos genéricos (como fórmulas de cortesía o siglas repetitivas) como palabras extremadamente comunes o raras. Este filtrado, junto a un listado ampliado de *stop words* específicas del dominio Enron, permitió generar un vocabulario relevante y manejable.

Este perfil temático no solo refleja la cultura comunicativa de cada grupo, sino que también sienta las bases para tareas como la clasificación automática de correos, detección de anomalías lingüísticas o diseño de encuestas adaptadas al lenguaje de cada equipo.

**6. Aplicación en negocio**

Ahora con todo este análisis y datos hemos puesto en práctica el trabajo anterior desarrollando ideas de negocio útiles, puestas en práctica con los correos de enron.

### **6. 1 Clasificación de correos por departamentos(Recruiter Fit)**

En esta segunda opción de Recruiter Fit avanzado, partimos de la premisa de que cada departamento —definido previamente como comunidad en el grafo de correos— posee un perfil lingüístico distintivo.  
 Para capturar ese estilo, se construyó un vectorizador TF-IDF que resume globalmente el uso del lenguaje en Enron. Luego, se agruparon estos vectores por cluster y se calculó el centroide de cada uno, es decir, su “huella léxica” promedio.

Al recibir un texto de candidato (correo de muestra o párrafo de prueba), se transforma con el mismo vectorizador y se calcula su similitud de coseno respecto a cada centroide. El resultado es un ranking de departamentos ordenados por cercanía semántica: los tres primeros son aquellos con mayor afinidad en temas, vocabulario y estilo.

Para enriquecer la recomendación, el sistema también extrae palabras clave compartidas entre el texto del aspirante y el vocabulario representativo de los tres clusters más cercanos. Así, además de saber que el “Cluster 1” obtuvo una similitud de 0.1430, se identifican hasta cinco términos coincidentes (por ejemplo: budget, proposal, revenue, analysis, expense) que explican el encaje.

Este enfoque ofrece ventajas prácticas:

* Alinea el estilo comunicativo de nuevos empleados con el tono y prioridades de cada equipo.  
  Facilita la rotación interna o asignación de proyectos, dirigiendo perfiles hacia el entorno donde aporten más rápidamente.
* Aporta transparencia al mostrar no solo la recomendación, sino también las palabras clave que la sustentan.
* Así, Recruiter Fit evoluciona de una simple etiqueta a una herramienta basada en análisis de redes, modelado del lenguaje y métricas semánticas que guía de forma inteligente la selección y asignación de talento.

### **6.2 Panel de visualizaciones del estado y evolución de la cultura empresarial**

Este caso de negocio busca ofrecer una visión clara y rápida de la cultura empresarial y su evolución, utilizando gráficos y tablas para revelar estilos comunicativos y actitudes internas. El objetivo es segmentar comportamientos, optimizar flujos de información y alinear mejor los equipos.

En un panel de cultura empresarial, se visualiza quién lidera la conversación, qué temas dominan y cómo se comparan los patrones de trabajo entre departamentos. Con base en los análisis anteriores, se construye la Tabla 5: Análisis por departamento.

La Figura 9 (Heatmap: TF-IDF promedio por departamento) ofrece una “radiografía” visual del uso lingüístico. Al representar con color la relevancia media de palabras clave, se detectan fácilmente picos léxicos distintivos:

* Cluster 0: términos eléctricos-industriales.
* Cluster 3: jerga transaccional (deals, sellers).
* Cluster 5: vocabulario técnico (permit, pipeline).
* Cluster 6: sin picos claros, lenguaje genérico o disperso.

Estas diferencias permiten anticipar temas, identificar prioridades y detectar vacíos de información especializada por grupo.

Del análisis emergen diez perfiles lingüísticos:

* Clusters 0 y 3: núcleo de trading eléctrico (operaciones o. transacciones).
* Cluster 1: finanzas corporativas.
* Cluster 2: cumplimiento regulatorio.
* Clusters 4 (RR.HH.), 7 (TI), 9 (documentación): soporte con vocabularios funcionales.
* Cluster 5: permisos de gas, infraestructura.
* Cluster 8: proyectos globales e internacionales.
* Cluster 6: grupo misceláneo, potencialmente spam.

Esta segmentación optimiza la comunicación, permite contenidos personalizados y favorece la integración entre áreas con afinidades léxicas, aislando flujos especializados para reducir ruido.

La Figura 10 (Dendrograma lingüístico por departamento) sintetiza la complejidad textual en una jerarquía clara de similitud semántica. Mientras los heatmaps muestran qué palabras destacan, el dendrograma revela quién se parece a quién, facilitando decisiones sobre fusión o separación de flujos comunicativos.

Cinco bloques principales se identifican:

* Clusters 8 + 5: proyectos y permisos internacionales.
* Clusters 2 + 0: operaciones y cumplimiento.
* Clusters 1 + 9: finanzas y documentación.
* Cluster 3: aislado por jerga de mercado y transacciones.
* Clusters de soporte (4, 7) y el misceláneo (6) requieren tratamiento específico.

La Figura 11 (Heatmap de prevalencia temática) traduce los resultados del modelo LDA en una matriz cromática, donde cada color representa el nivel de atención que cada departamento dedica a los distintos temas (T0–T3). Esta visualización permite identificar rápidamente las áreas de enfoque de cada grupo:

* T0 (Noticias corporativas): domina en casi toda la red, con el Cluster 4 como principal emisor.
* T1 (Planificación y gestión): aparece con intensidad moderada en varios departamentos, reflejando tareas operativas y de coordinación.
* T2 (Análisis de mercado): tiene baja presencia general, lo que sugiere la necesidad de un boletín especializado para insights financieros.
* T3 (Logística de reuniones): destaca en el Cluster 2, que podría beneficiarse de herramientas de automatización de agendas o resúmenes de minutas.

Resumen por Cluster

* Clusters 0 y 2: enfoque en noticias y planificación, eje Operaciones–Cumplimiento.
* Cluster 1: finanzas + análisis de mercado.
* Cluster 3: transacciones y datos bursátiles.
* Cluster 4: canal central de anuncios.
* Cluster 5: noticias + permisos regulatorios.
* Clusters 6, 7 y 8: distribuciones planas, áreas transversales o internacionales.

Este heatmap temático facilita la creación de feeds informativos específicos, automatización de agendas y alineación de contenidos según necesidades reales.

Los hallazgos demuestran que la cultura empresarial se puede cartografiar a través del lenguaje. Proponemos una plataforma de inteligencia que:

* Diagnostique el ADN léxico de cada unidad con heatmaps y dendrogramas.
* Segmenta flujos de información, enviando noticias o alertas solo a los equipos relevantes.
* Monitoree la evolución cultural, permitiendo ver cambios tras fusiones, crisis o nuevas políticas.
* Escale fácilmente a cualquier organización con archivos de correos o chats, ofreciendo un dashboard plug-and-play con insights sobre cultura, eficiencia y riesgo.

Los tres paneles demuestran que el lenguaje corporativo es un activo medible y estratégico. Analizar los correos internos permite detectar estilos de comunicación, prioridades temáticas y dinámicas organizativas con precisión.

Convertir este lenguaje en inteligencia accionable mejora la toma de decisiones, optimiza la asignación de talento y facilita la integración entre áreas afines. Además, permite monitorear la evolución cultural tras cambios clave y detectar sinergias o vacíos informativos.

La metodología es escalable y replicable: cualquier empresa con datos de comunicación puede implementar un sistema que traduce texto en insights operativos, culturales y estratégicos.

En síntesis, este enfoque convierte la comunicación interna en una herramienta clave para diseñar organizaciones más eficientes y conectadas a la cultura empresarial.

# **Inferencia de urgencia e importancia de los correos electrónicos de Enron**

## **1.Carga y limpieza del dataset**

Se carga el conjunto de correos desde el archivo CSV ya parseado y limpiado como previamente se ha explicado. Se implementan controles de calidad y tolerancia ante errores de formato, y se eliminan los correos vacíos o malformados.

## **2. Enriquecimiento temporal y estructural.**

Con el objetivo de incorporar contexto al contenido textual del correo, se crean variables temporales:

* Franja horaria (franja\_horaria): clasifica cada correo como enviado de madrugada, mañana, tarde o noche, a partir del campo hora (decimal).
* Día de la semana (dia\_semana): convierte desde el campo fecha, para detectar patrones asociados a días laborales versus fines de semana.
* Indicador de fin de semana (es\_finde): booleano, útil para identificar correos enviados en momentos no laborables.
* Indicador de horario laboral (es\_laboral): correos enviados de lunes a viernes entre las 9:00 y 18:00.
* Número de destinatarios (n\_destinatarios): conteo del total de correos en el campo to, útil como indicador de correos masivos o informativos.

Estas variables estructurales serán utilizadas posteriormente como entrada junto con los embeddings semánticos del texto, permitiendo al modelo incorporar no solo el contenido del mensaje sino también el contexto situacional en el que fue enviado.

## **3. Etiquetado heurístico inicial**

Dado que el conjunto de datos no incluía etiquetas explícitas de urgencia, se define una estrategia heurística para clasificar los correos en cuatro categorías: *urgente*, *importante*, *rutina* y *prescindible*. Esta heurística se basa en la presencia de palabras clave en el cuerpo y asunto del mensaje (como “urgent”, “asap” o “please check”), en combinación con factores contextuales como el horario de envío (laboral o no laboral) y el número de destinatarios. Se prioriza una alta cobertura y un sesgo conservador para garantizar que los correos potencialmente urgentes no sean subestimados. (Figura 12: Distribución Porcentual de Etiquetas de Urgencia)

Estas etiquetas se utilizan como referencia durante el entrenamiento supervisado del modelo y como base para la evaluación inicial del rendimiento predictivo.

## **4. Extracción de embeddings semánticos con DistilBERT**

Con el objetivo de capturar la representación semántica de cada mensaje, se utiliza el modelo preentrenado DistilBERT de Hugging Face. Este modelo, una versión reducida y eficiente de BERT, permite extraer embeddings contextuales a nivel de frase.

Cada correo se vectoriza a partir del campo clean\_body. Se tokeniza el texto y se obtiene el embedding de la secuencia utilizando la salida del [CLS] token, correspondiente a la representación agregada de la frase completa. Para ello se emplea el modelo distilbert-base-uncased, sin afinamiento adicional (feature extraction mode), lo que reduce el riesgo de sobreajuste y facilita la integración con modelos clásicos.

Los embeddings generados (de dimensión 768) se concatenan posteriormente con las variables estructurales extraídas (horario, destinatarios, etc.), formando una matriz de entrada enriquecida para los clasificadores supervisados. Este enfoque permite aprovechar tanto la semántica del contenido como el contexto situacional del mensaje.

## **5. Preparación del conjunto de entrenamiento y holdout, y normalización y combinación de features**

Se separa un conjunto holdout (10%) estratificado por etiqueta para evaluación final. El 90% restante se utiliza para entrenamiento y validación. Los embeddings de DistilBERT se combinan con las variables estructurales normalizadas, formando la matriz final de entrada para el modelo.

## **6. Entrenamiento de modelos**

### **Random Forest como baseline interpretativo**

El primer modelo entrenado fue un Random Forest Classifier, utilizado como baseline por su robustez, capacidad de manejo de datos tabulares heterogéneos y facilidad de interpretación de importancia de variables. Se utilizó el conjunto df\_model, que contiene tanto los embeddings generados con DistilBERT como las variables estructurales enriquecidas (franja horaria, número de destinatarios, horario laboral, entre otras).

La división del dataset fue del 90% para entrenamiento y 10% para test, aplicando además una validación cruzada estratificada de 5 particiones (Stratified K-Fold CV) para estimar la generalización del modelo. Las métricas obtenidas con Random Forest mostraron un rendimiento moderado:

* F1-score macro en torno al 0.70
* Precisión elevada en clases frecuentes como *importante*
* Sin embargo, una recall baja en la clase *urgente*, sugiriendo que el modelo no logra capturar adecuadamente los patrones menos explícitos asociados a esa clase. Este comportamiento es coherente con las limitaciones conocidas de Random Forest en tareas con clases desbalanceadas y presencia de alta correlación entre variables derivadas (por ejemplo, embeddings semánticos).

A pesar de sus limitaciones, este modelo resulta útil como instrumento de interpretación, especialmente al analizar la importancia de las variables estructurales, donde variables como hora, es\_laboral o n\_destinatarios presentaron relevancia en las decisiones del modelo. La figura 13: Matriz de Confusión de Random Forest, confirma que rinde bien en importante y prescindible. Sin embargo, la clase urgente sigue siendo confundida con importante (26 veces), lo que indica que aún hay margen de mejora en su detección.

### **XGBoost como modelo principal**

A continuación, se entrenó un modelo XGBoost Classifier como solución principal. XGBoost ha demostrado ser altamente competitivo en tareas supervisadas con clases múltiples, especialmente al manejar relaciones no lineales y regularización automática.

Se utilizó el objetivo multi:softmax con num\_class=4 para clasificación directa de las cuatro categorías: *urgente*, *importante*, *rutina* y *prescindible*. Las etiquetas fueron codificadas con LabelEncoder, y la evaluación se realizó tanto en el conjunto de test como mediante validación cruzada estratificada (5-fold), replicando el pipeline de Random Forest para comparación equitativa.

Los resultados de XGBoost mostraron una mejora significativa en comparación con el baseline:

* F1-score macro cercano a 0.77
* Mejora destacada en la clase *urgente*, con un incremento de ~15 puntos porcentuales en recall
* Reducción del error sistemático de clasificación de *urgente* como *prescindible* o *importante*
* Mayor estabilidad en clases minoritarias gracias al uso combinado de embeddings y variables estructurales.

Además, la matriz de confusión (Figura 14: Matriz de Confusión de XGBoost) evidenció una mejor diferenciación entre las clases, con especial énfasis en las fronteras más borrosas como *importante* vs *rutina* y *urgente* vs *importante*. Este comportamiento puede atribuirse a la capacidad de XGBoost de explotar interacciones entre los componentes semánticos (embeddings) y variables situacionales (horarios, destinatarios, etc.), lo que otorga al modelo mayor contexto al momento de clasificar.

## **7. Validación cruzada y evaluación clásica. Análisis de errores críticos y léxico**

Se implementó una validación cruzada estratificada (StratifiedKFold, k=5) para estimar la capacidad de generalización del modelo y evitar sobreajuste, empleando las mismas particiones en los modelos de Random Forest y XGBoost. En paralelo, se realizó una evaluación clásica sobre el conjunto de test (10%), utilizando métricas estándar como precision, recall y F1-score macro. Los resultados confirmaron la superioridad de XGBoost, particularmente en clases minoritarias como *urgente*, donde mejoró significativamente el recall.

Posteriormente, se llevó a cabo un análisis manual de errores más críticos, centrado en falsos negativos de la clase *urgente*. En dichos casos, se observaron patrones léxicos específicos como “immediately”, “please respond” o estructuras de frase que denotaban urgencia implícita. Esta revisión permitió identificar lagunas en la heurística original y posibles líneas para refinar el conjunto de entrenamiento o añadir reglas adicionales como features. El análisis también confirmó que errores en la clase *rutina* estaban asociados a correos informales o reenviados sin contenido sustantivo, lo cual sugiere ambigüedad estructural más que fallo del modelo.

## **8. Evaluación Final del Modelo sobre el Conjunto de Validación Externa (Holdout)**

La evaluación final del modelo se ha llevado a cabo utilizando un conjunto de validación externa (holdout), correspondiente al 10 % de la muestra total. Este conjunto no ha sido utilizado ni en el entrenamiento del modelo ni en la fase de validación cruzada, lo que permite obtener una estimación robusta de la capacidad de generalización del clasificador. En la Tabla 6: Evaluación por clase frente a holdout set, podemos observar los resultados.

Estos resultados alcanzados han sido notablemente sólidos. El modelo obtuvo una precisión global (accuracy) del 87 %, un F1-score macro promedio de 0.869 y un F1-score ponderado de 0.867. Estos valores indican no sólo un alto rendimiento general, sino también una adecuada distribución del rendimiento entre las distintas clases, lo cual es especialmente *importante* en contextos multiclase con distribución potencialmente desbalanceada.

Al desglosar los resultados por clase, se observa lo siguiente:

* La clase *prescindible* se predice con gran fiabilidad (F1 = 0.920, recall = 0.986), lo cual sugiere que el modelo ha aprendido a identificar patrones asociados a correos de baja prioridad con alta precisión.
* La clase *importante* presenta también una elevada capacidad de predicción (F1 = 0.814), aunque muestra una ligera tendencia a ser confundida con otras clases, particularmente con *prescindible* y *rutina*.
* En el caso de la clase *rutina*, se alcanzó un F1-score de 0.885, con un recall de 0.964, lo cual refleja un excelente comportamiento del modelo para identificar comunicaciones repetitivas o sistemáticas.
* Por último, la clase *urgente* fue clasificada con un F1-score de 0.857 y una precisión perfecta (1.000), aunque con un recall del 75 %. Esto indica que, si bien el modelo es muy conservador a la hora de etiquetar un correo como *urgente*, puede dejar sin clasificar correctamente algunos correos realmente urgentes, clasificándolos mayoritariamente como *importantes*.

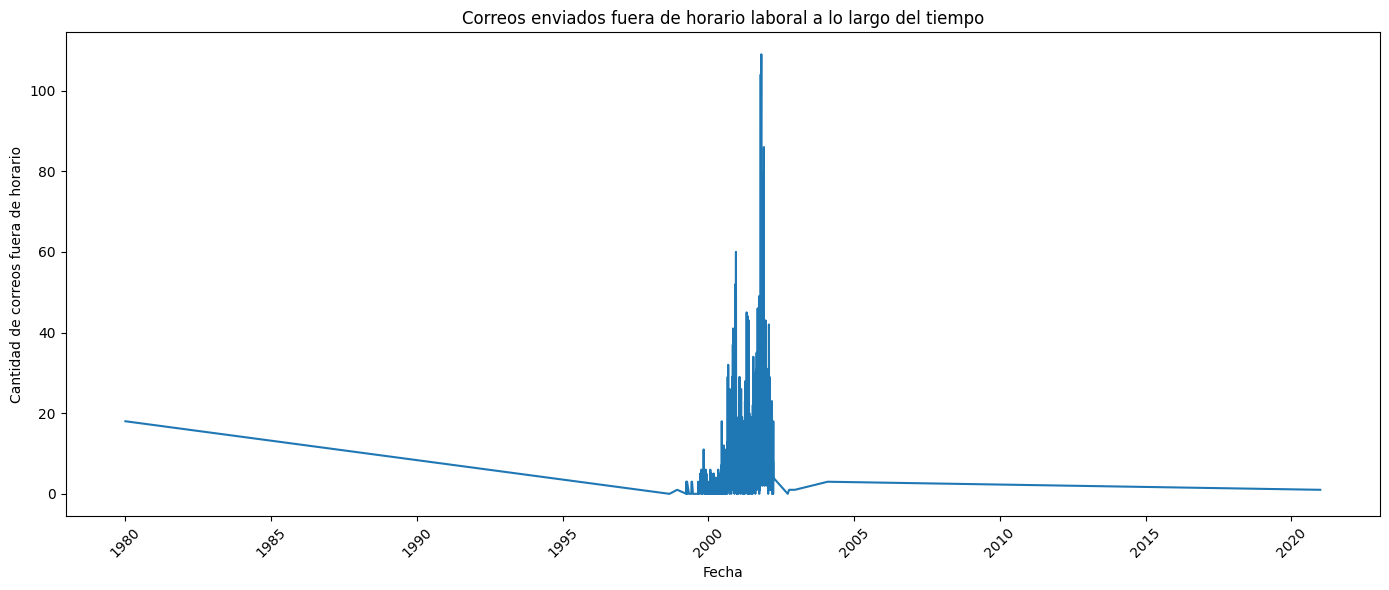
La matriz de confusión (Figura 15: Matriz de Confusión de Validación Externa) confirma este patrón: los errores más significativos se concentran en la confusión entre las clases *urgente* e *importante*, un comportamiento esperable dadas sus similitudes semánticas. Es preferible, desde una perspectiva operativa, que el modelo etiquete un correo *urgente* como *importante* antes que como *prescindible* o de *rutina*, lo cual se cumple satisfactoriamente.

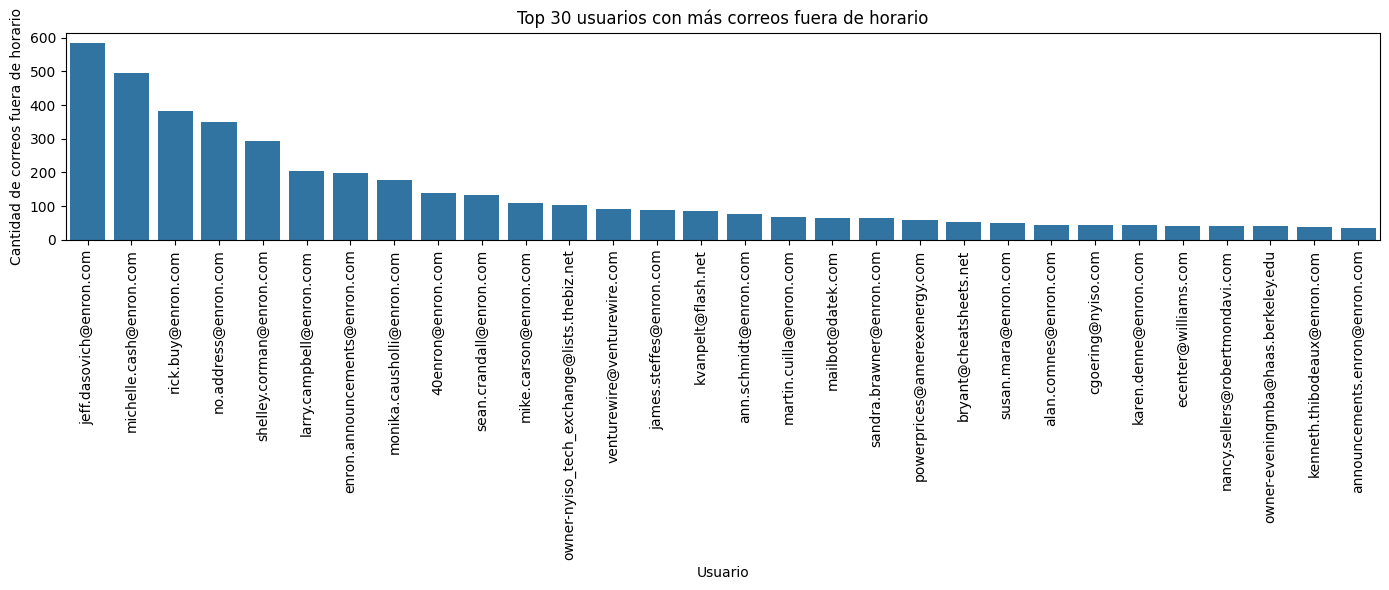
En conjunto, el modelo demuestra una alta capacidad de generalización y un balance adecuado entre precisión y exhaustividad en las cuatro clases de urgencia. El pipeline desarrollado —que integra variables estructurales, representaciones textuales mediante embeddings, y un clasificador XGBoost— ha demostrado ser efectivo para la tarea planteada. Se recomienda, si se dispone de más tiempo, explorar estrategias como ajuste de umbrales de decisión o penalización de errores tipo II, para mejorar la sensibilidad del modelo en la detección de comunicaciones urgentes.

## **9. Aplicación empresarial del modelo de inferencia de urgencia en correos electrónicos**

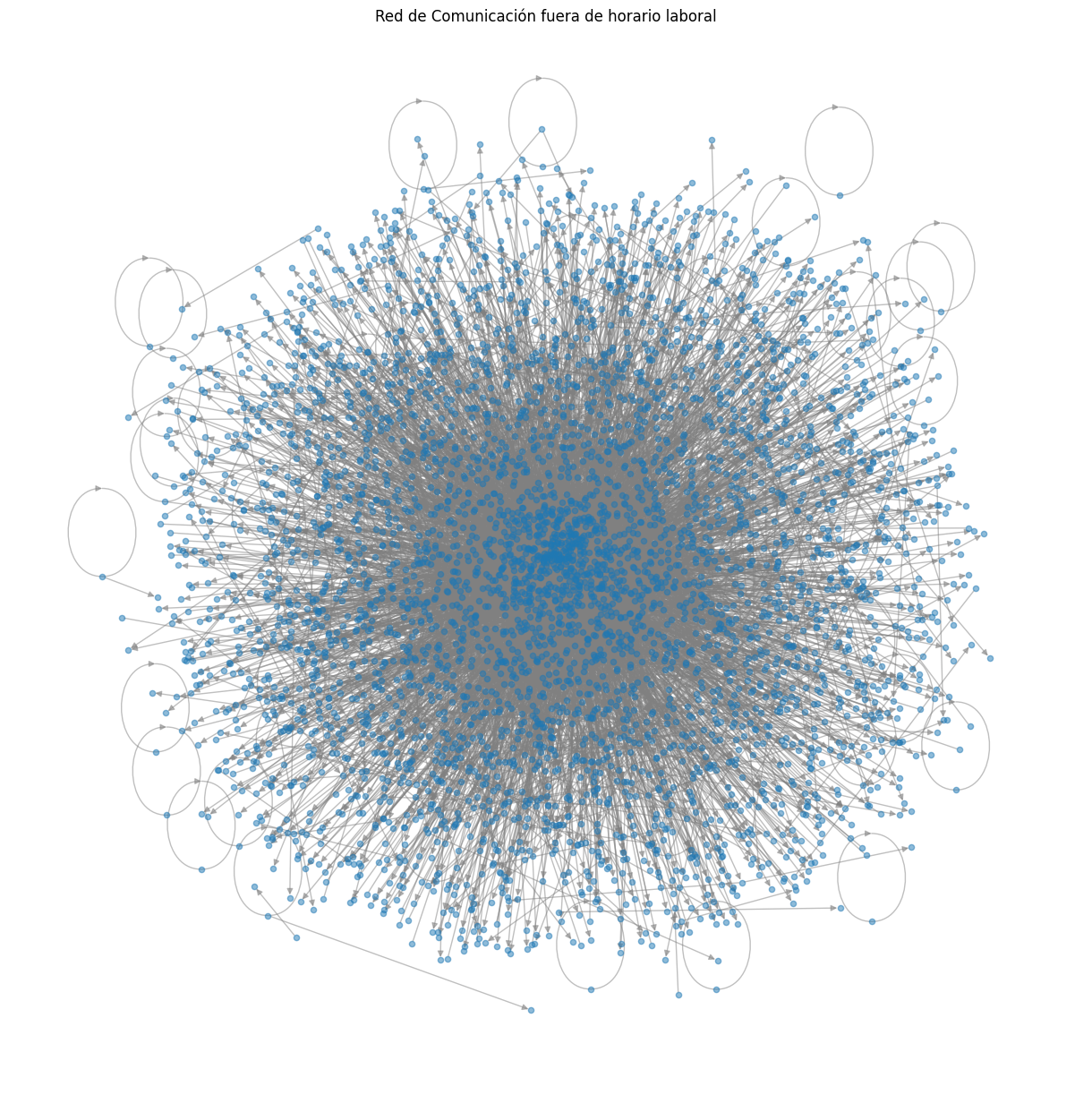
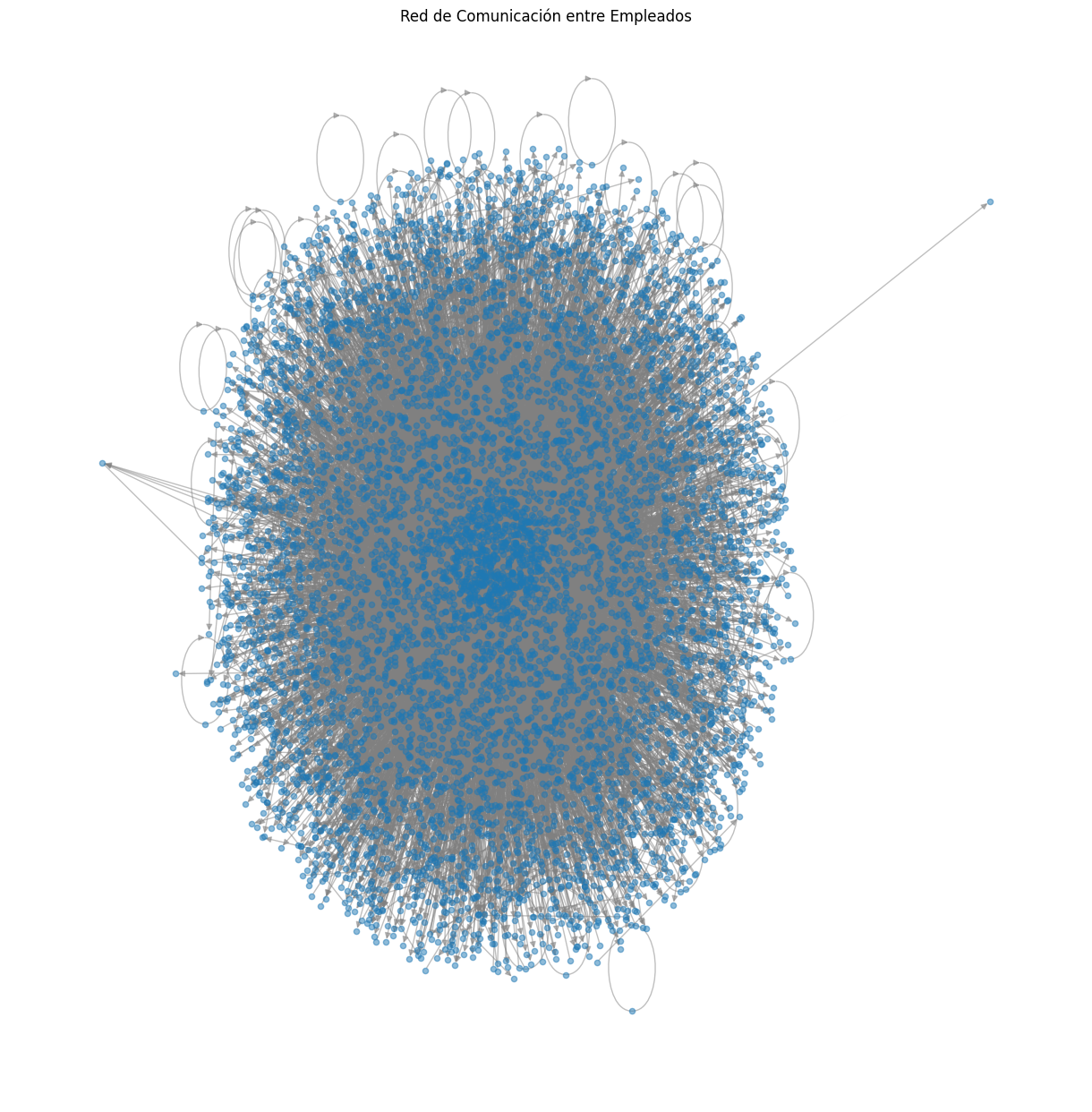
1. Priorización automática del flujo de trabajo: Ordenación dinámica de bandejas de entrada, enfocando la atención en correos con alto impacto temporal. Alertas tempranas ante mensajes que requieren acción inmediata, especialmente útil para roles ejecutivos o de operaciones críticas. Delegación automatizada, redirigiendo correos prescindibles o rutinarios a asistentes o sistemas de gestión documental.
2. Reducción de la sobrecarga de información, filtrado inteligente de mensajes prescindibles o repetitivos: Liberar tiempo de los empleados. Disminuir la fatiga por exceso de correos (email fatigue). Aumentar la concentración en tareas de alto valor añadido.
3. Soporte en entornos de compliance o riesgo operacional, detección de mensajes urgentes fuera del horario laboral: Identificar patrones de estrés laboral o posibles incumplimientos normativos. Monitorizar comunicaciones clave durante eventos críticos o incidentes técnicos.
4. Análisis de cultura organizacional y comunicación interna: Departamentos con tendencia a la escalada innecesaria de urgencia. Horarios o días con sobrecarga comunicacional. Patrones de comunicación informal vs formal que impactan la productividad.

# **Anexo**

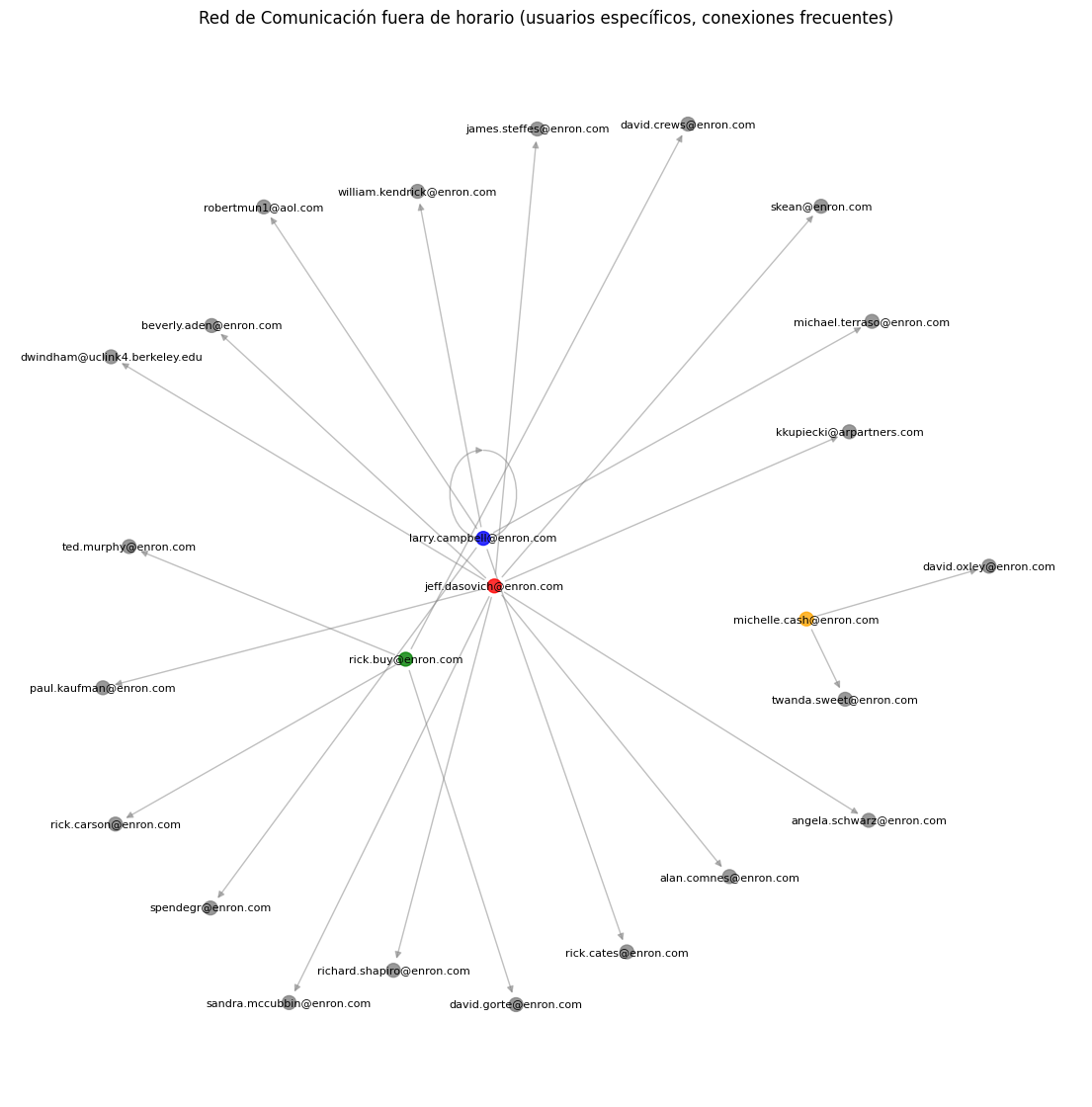
*Figura 1: Número de emails enviados fuera del horario laboral a lo largo del tiempo.*



*Figura 2: Frecuencia de emails fuera de horario laboral de los 30 empleados que más emails enviaron.*



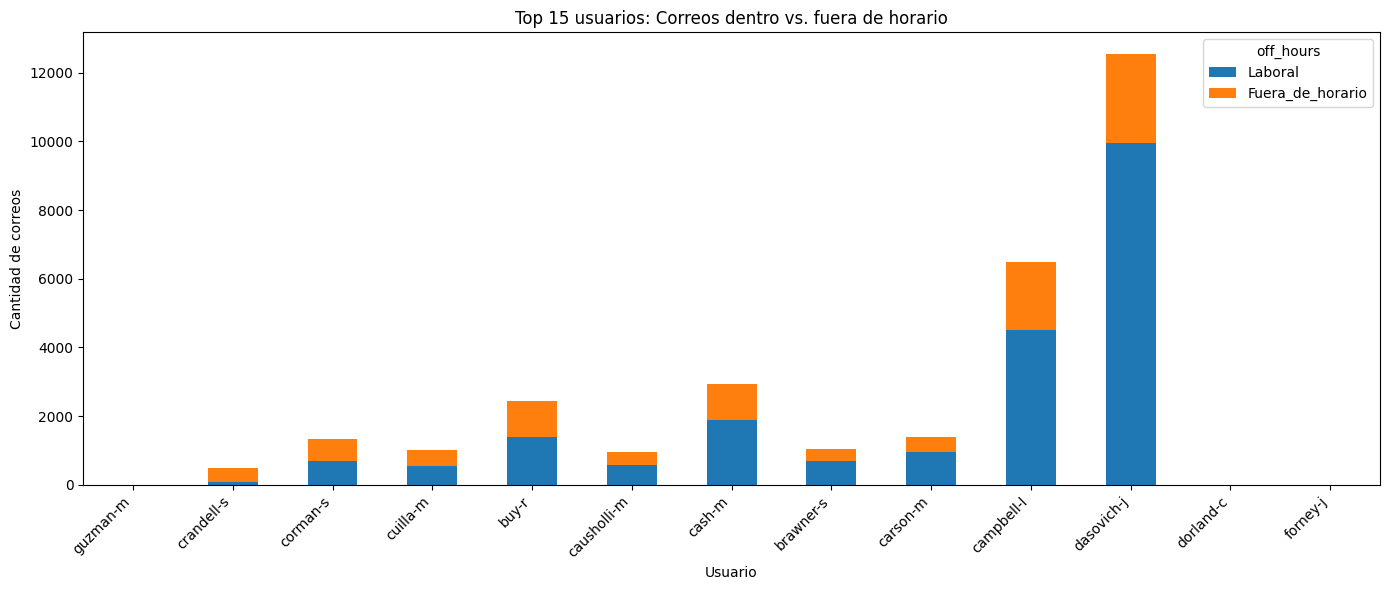
*Figura 3: Red de comunicación entre empleados (izquierda) y red de comunicación entre empleados fuera del horario laboral (derecha).*



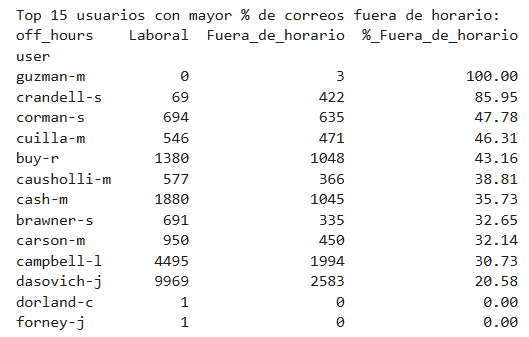
*Figura 4: Red de comunicación fuera del horario laboral de empleados específicos.*



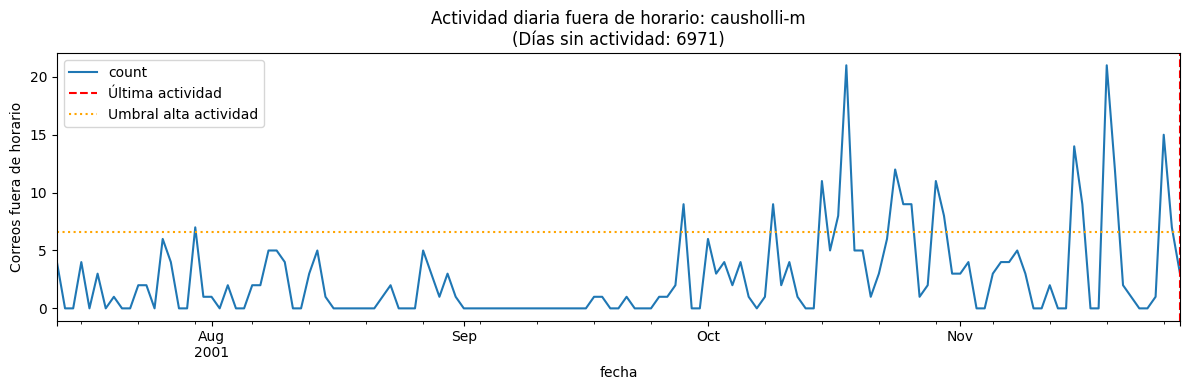
*Tabla 1: Métricas de red de comunicación entre empleados específicos fuera de horario laboral.*

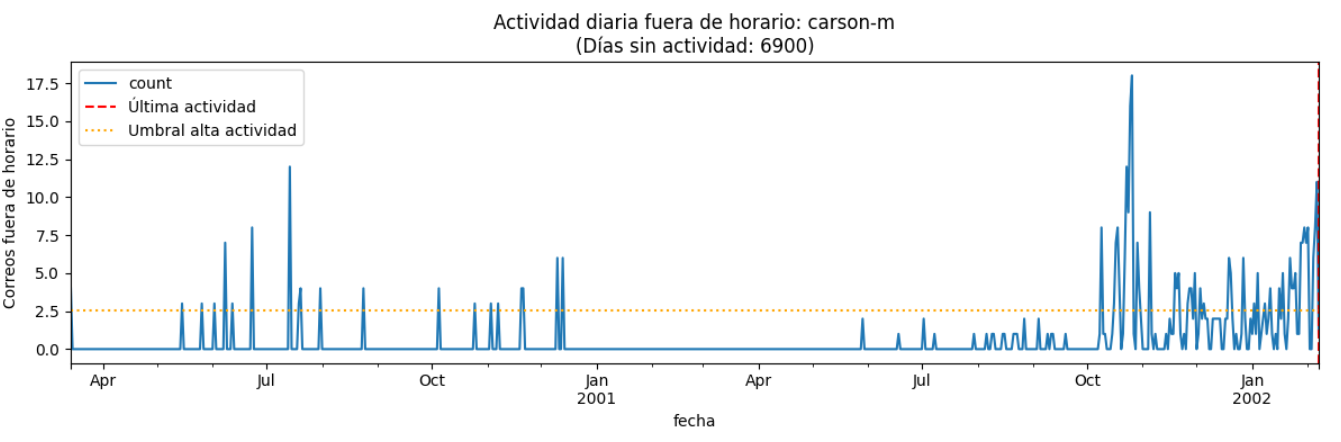


*Figura 5: Top 15 empleados con mayor porcentaje de emails fuera (naranja) de horario laboral*

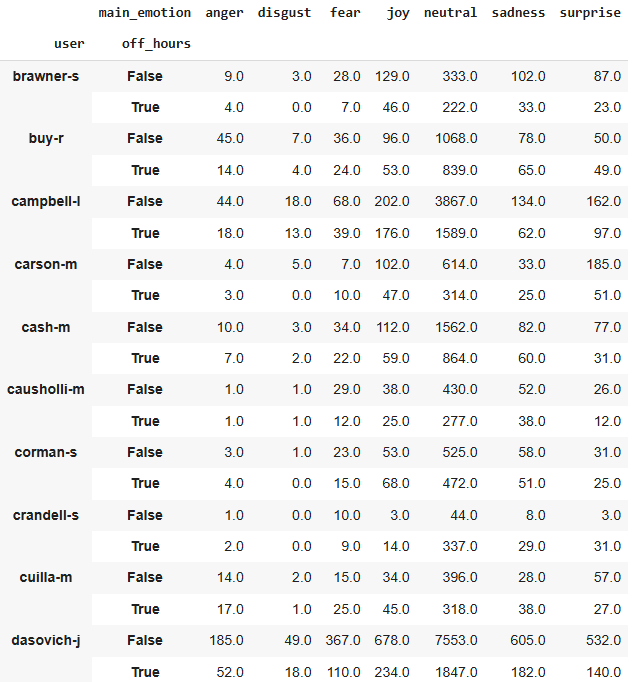


*Tabla 2: Top 15 empleados con mayor porcentaje de emails fuera del horario laboral*

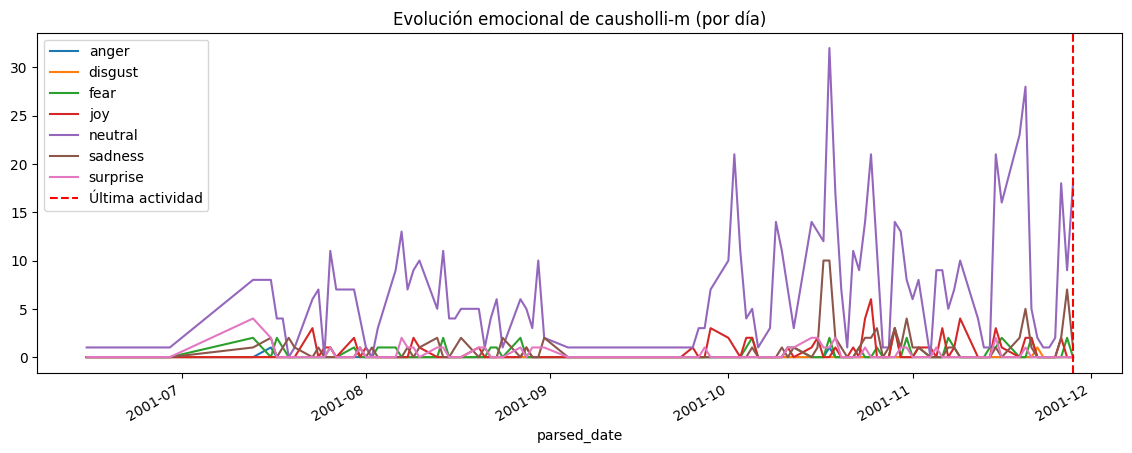




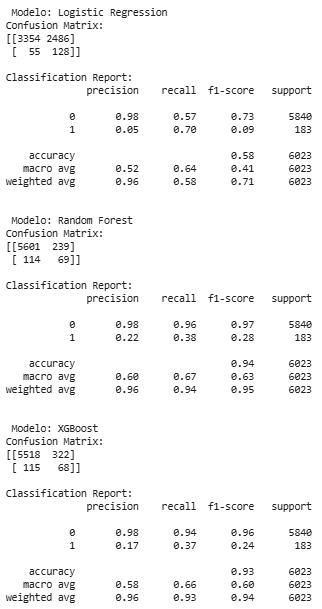
*Figura 6: Ejemplos de actividad diaria enviando emails de trabajo fuera de horario (causholli-m y carson-m).*

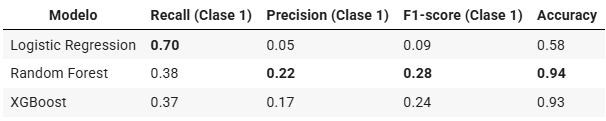
**

*Tabla 3: Clasificación por emociones en horario laboral vs. fuera de horario.*

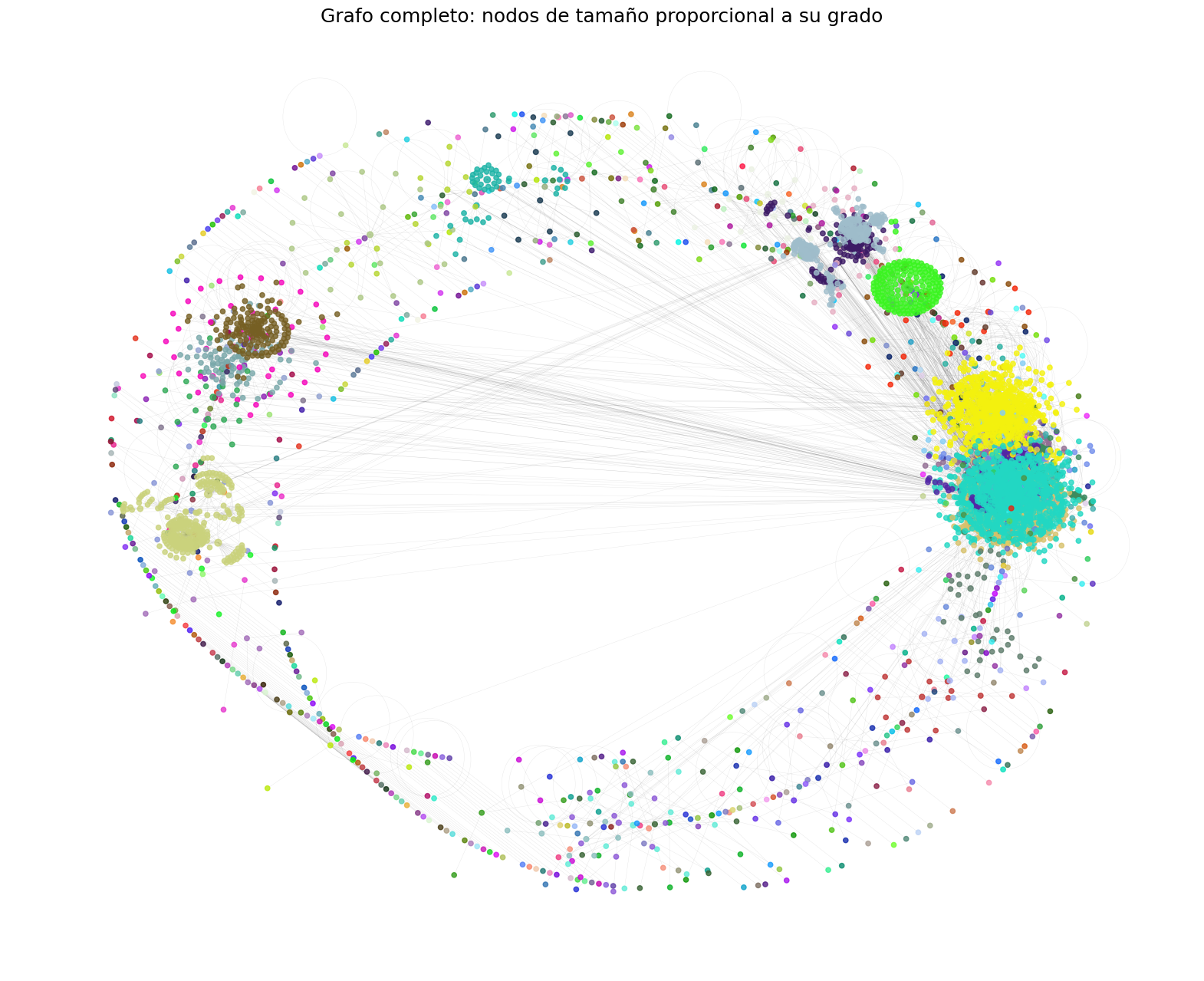


*Figura 7: Ejemplo de la evolución emocional de un empleado (causholli-m) justo antes de desaparecer.*

**

**

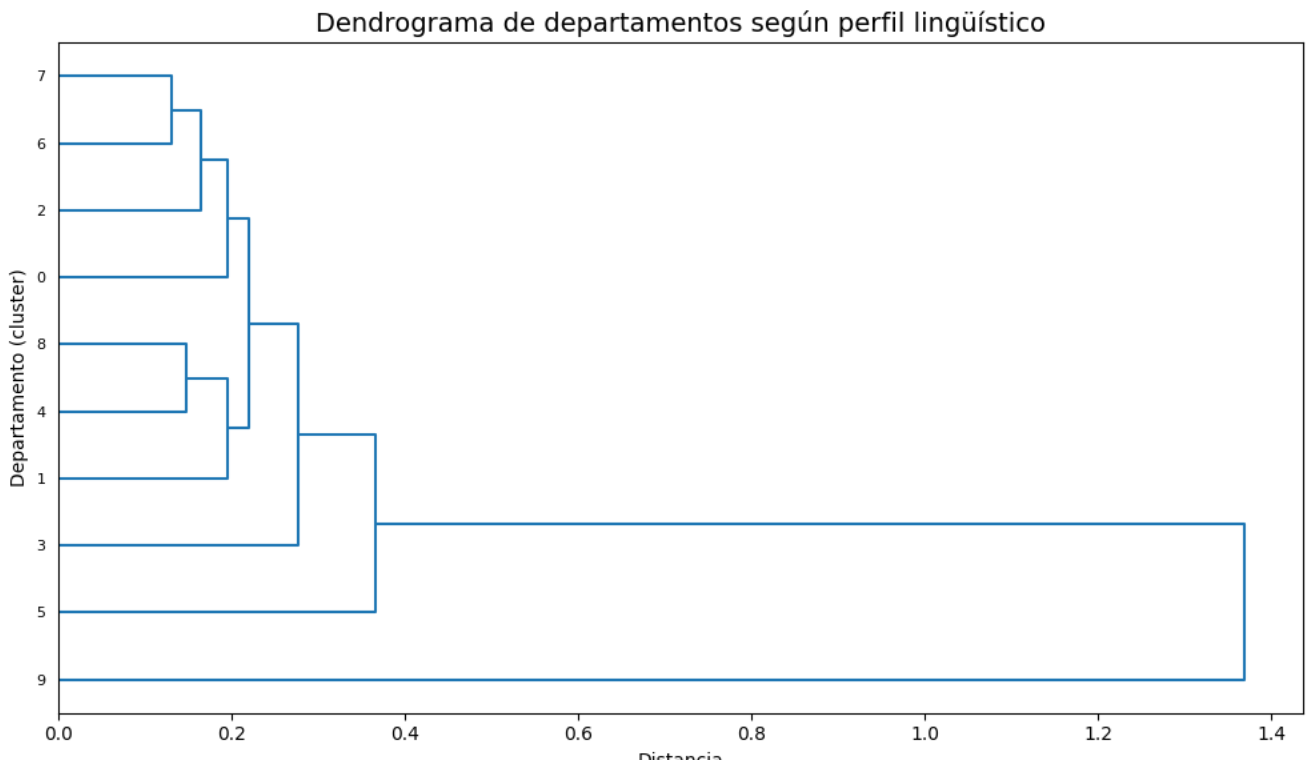
*Tabla 4: Métricas de las predicciones de los distintos modelos (ver ML\_burnout.ipynb).*

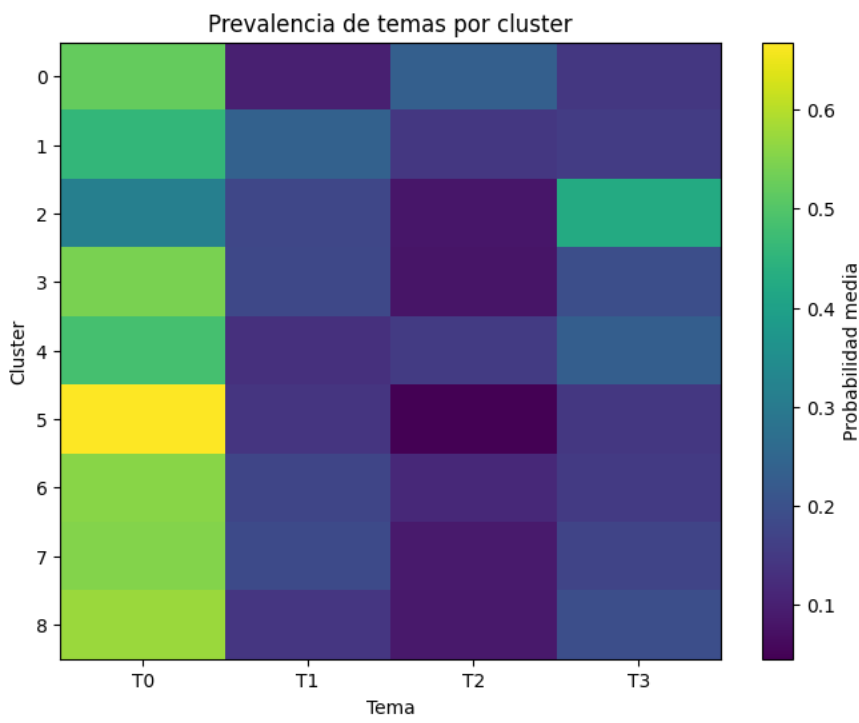
  
*Figura 8: Grafo de departamentos por comunicación.*

| **departamento\_cluster** | **hora\_promedio\_envio** | **porcentaje\_laboral** | **destinatarios\_promedio** | **total\_correos** | **top\_keywords** |
| --- | --- | --- | --- | --- | --- |
| 0 | 14.103565 | 77.979605 | 2.13464 | 12552 | sellers, edison, dasovichnaenron, haas, electric... |
| 1 | 16.809615 | 51.961824 | 1.582185 | 943 | causholli, pulp, paper, hello, industrial, division... |
| 2 | 13.961471 | 69.086146 | 3.301741 | 6489 | permit, pcb, campbell, pjm, datek, transwestern... |
| 3 | 15.467415 | 56.383855 | 1.949341 | 2428 | rac, var, london, ted, published, named, executive... |
| 4 | 13.406789 | 11.405295 | 1.523422 | 491 | sean, bpa, mc, product, west, mw, rto, deals... |
| 5 | 15.119926 | 61.811966 | 1.379145 | 2925 | cashhouectect, hr, cashenron, employment, cash... |
| 6 | 14.188405 | 67.5 | 7.884286 | 1400 | mc, saturday, rb, yahoo, game, updated, k, bank... |
| 7 | 14.092298 | 58.581907 | 1.43912 | 2045 | martin, love, requests, visit, anderson, game... |
| 8 | 16.524003 | 47.479308 | 1.486832 | 1329 | corman, wireless, tw, kim, lynn, travel, pipeline... |

| **departamento\_cluster** | **top\_pagerank** | **top\_grado** | **top\_betweenness** |
| --- | --- | --- | --- |
| 0 | [jeff.dasovich@enron.com, susan.mara@enron.com...] | [jeff.dasovich@enron.com, susan.mara@enron.com...] | [jeff.dasovich@enron.com, mday@gmssr.com, mart...] |
| 1 | [,michelle.cash@enron.com, rick.buy@enron.com] | [,michelle.cash@enron.com, rick.buy@enron.com] | [,michelle.cash@enron.com, rick.buy@enron.com] |
| 2 | [larry.campbell@enron.com, shelley.corman@enron.com...] | [larry.campbell@enron.com, shelley.corman@enron.com...] | [shelley.corman@enron.com, larry.campbell@enron.com...] |
| 3 | [jdasovic@enron.com, nancy.sellers@robertmondavi.com...] | [jdasovic@enron.com, nancy.sellers@robertmondavi.com...] | [jdasovic@enron.com, csilva@redsky.com, nancy...] |
| 4 | [jae.black@enron.com, tom.hoatson@enron.com, c...] | [jae.black@enron.com, tom.hoatson@enron.com, c...] | [jae.black@enron.com, tom.hoatson@enron.com, C...] |
| 5 | [mike.carson@enron.com, no.address@enron.com...] | [mike.carson@enron.com, no.address@enron.com...] | [mike.carson@enron.com, powerprices@amerexener...] |
| 6 | [socrskibum@aol.com, ksu93dlv@aol.com, chjohns...] | [socrskibum@aol.com, ksu93dlv@aol.com, kmn30@h...] | [m\_besch@yahoo.com, socrskibum@aol.com, ksu93d...] |
| 7 | [lcampbel@enron.com, scorman@enron.com, mailbox...] | [lcampbel@enron.com, scorman@enron.com, mcuill...] | [lcampbel@enron.com, scorman@enron.com, mcuill...] |
| 8 | [jblunden@xenergy.com, jwu@apx.com, anders.gla...] | [jblunden@xenergy.com, jwu@apx.com, jweisgall@...] | [jblunden@xenergy.com, vjw@ceert.org, jwu@apx...] |

*Tabla 5: Tabla de análisis por departamento.*

  
*Figura 9: Heatmap de palabras clave*  
*Figura 10: dendrograma de departamentos según su perfil lingüístico.*

  
*Figura 11: Heatmap topics.*

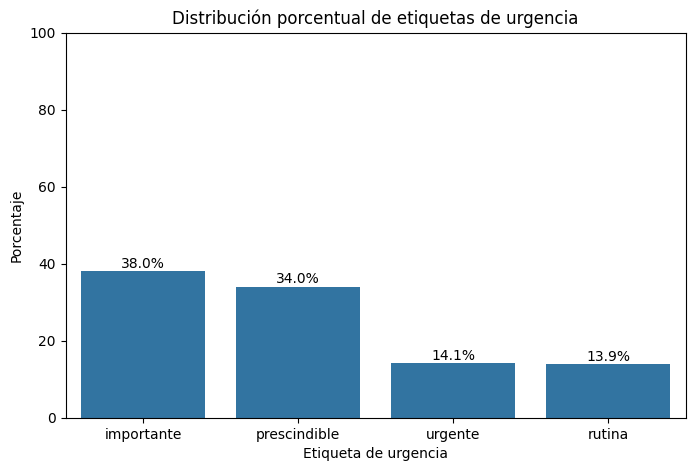
**

Figura 12: Distribución Porcentual de Etiquetas de Urgencia

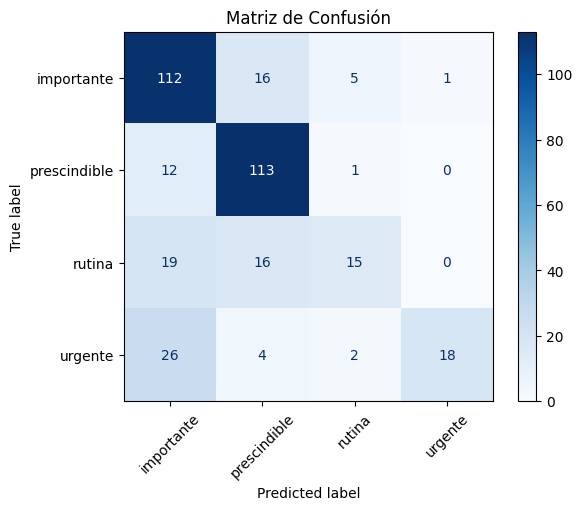


Figura 13: Matriz de Confusión de Random Forest

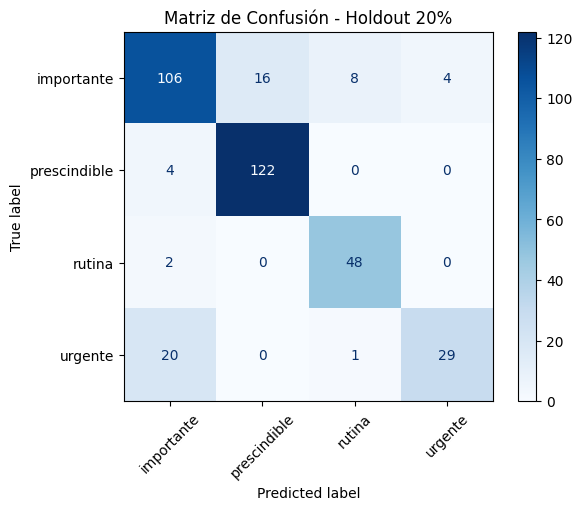


Figura 14: Matriz de Confusión de XGBoost

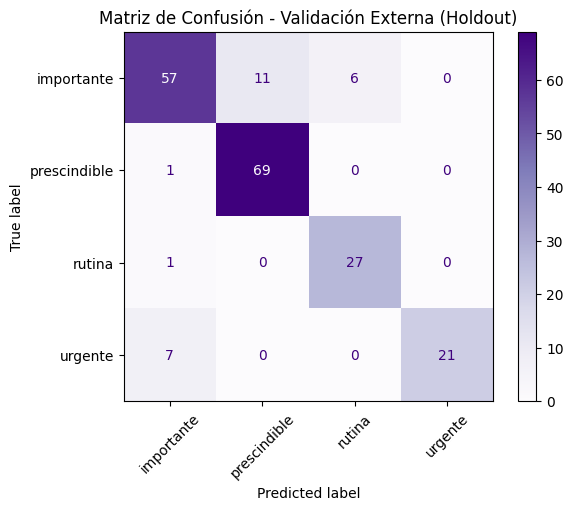


Figura 15: Matriz de Confusión de Validación Externa

| ***Clase*** | ***Precision*** | ***Recall*** | ***F1*** | ***Observaciones*** |
| --- | --- | --- | --- | --- |
| ***importante*** | *0.864* | *0.770* | *0.814* | *Algo de confusión con prescindible y rutina, pero en general buen equilibrio.* |
| ***prescindible*** | *0.863* | *0.986* | *0.920* | *Excelente recall, casi todos los prescindibles se predicen correctamente.* |
| ***rutina*** | *0.818* | *0.964* | *0.885* | *Fuerte y precisa, aunque puede estar capturando parte de lo urgente o importante.* |
| ***urgente*** | *1.000* | *0.750* | *0.857* | ***Muy alta precisión****, pero se le escapan algunos (7 se predicen como importante).* |

*Tabla 6: Evaluación por clase frente a holdout set*