



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



ESCUELA TÉCNICA  
SUPERIOR INGENIERÍA  
INDUSTRIAL VALENCIA

**TRABAJO FIN DE MASTER EN INGENIERÍA INDUSTRIAL**

# **DISEÑO E IMPLANTACIÓN DE UNA METODOLOGÍA Y HERRAMIENTA PARA MEJORAR LA DETECCIÓN Y EL ANÁLISIS DE TENDENCIAS EN FACTORES DE RIESGOS NO FINANCIEROS**

AUTOR: Vincent Thomas Rizzo

TUTOR: Julien Maheut

Selección

**Curso Académico: 2019-20**



# Agradecimientos

Gracias a mi familia, primero, por darme la oportunidad de realizar la carrera que me apasiona, por apoyarme financiera y psicológicamente durante estos 7 años. Me habéis enseñado la disciplina y la humildad que me permitirán realizar mis sueños y aportar a los demás.

A mis compañeros de Francia, de España, Italia, México u otros países, gracias por compartir vuestra cultura conmigo, aceptar la mía en el respeto mutuo que se debe, gracias a vosotros aprendí mucho tanto sobre el mundo como sobre mi mismo.

A mis colegas de Datamaran, gracias por dejarme formar parte del viaje con vosotros, por siempre estar atentos y pacientes cuando tenía que ponerme al día con el funcionamiento de la empresa y de las herramientas. Mención especial a Jeson, mi mentor, gracias a quién el proyecto de este trabajo ha sido posible, y a Jérôme, mi tutor, por haberme dado la oportunidad de unirme al equipo.

Gracias a la escuela CentraleSupélec por los años que pasé allá y por la posibilidad de realizar el intercambio que quería hacer. Gracias a la Universitat Politècnica de València, por acogerme los últimos 2 años y especialmente la Escuela Técnica Superior de Ingeniería Industrial, por los conocimientos aprendidos. Si la formación que recibí no cuadra exactamente con el trabajo presente (que pertenece más a informática y matemáticas), la asignatura de Métodos Cuantitativos para la Organización Industrial ha sido sin duda muy útil para realizar el trabajo. También, al ser generalista, agradezco a la escuela por dejar libres los estudiantes de realizar un trabajo diferente de los demás para incentivar la creatividad en los trabajos. Gracias a todos el equipo de profesores y la administración de la escuela, especialmente mi tutor Julien Maheut, por ser siempre disponible y reactivo, y por sus consejos claros y útiles durante todo el ciclo de vida del trabajo.

Por fin, gracias a todas las personas que no he mencionado pero que me ayudaron a lograr mis objetivos, y acabar este trabajo.





# Resumen

Datamaran es una empresa creada en 2014 en Londres, con el objetivo de facilitar a las grandes empresas una herramienta innovadora que permite leer automáticamente y en tiempo real los documentos útiles para tomar decisiones estratégicas en el contexto medioambiental y social. Su departamento técnico se basa casi exclusivamente en Valencia, donde trabajan alrededor de 20 personas, desarrolladores y data scientists por la mayoría. El resto de sus empleados (50 personas en total) se reparte entre las oficinas de Londres y Nueva York, donde se encargan de las ventas, comunicación, asociaciones y atención al cliente. Dentro de los clientes de Datamaran se encuentran la BBVA, el banco Santander, Rakuten, Amazon, Philips y Morgan Stanley.

El trabajo se lleva sobre un módulo específico de la solución de Datamaran: el *continuous monitoring*. Este módulo se encarga de observar cómo evolucionaron los datos en los últimos 6 meses, para que los clientes puedan entender mejor cuáles son las posibles razones que explican por qué una temática dada se encuentra más importante hoy en día. Se llevará un diálogo con el departamento business de Datamaran para definir correctamente los *insights* interesantes para los clientes, se definirán indicadores medibles que permitan identificar tendencias en los datos e intentar predecir que temáticas más probablemente serán importantes en el futuro.

El documento se dividirá en 7 capítulos, después de la introducción, en el capítulo 2 se tratará del entorno del problema, en la empresa, poniendo el módulo en su contexto en la aplicación. Luego, en el capítulo 3 se llevará a cabo un análisis detallado del funcionamiento actual del módulo existente y en el capítulo 4, gracias al departamento responsable del producto se determinarán las especificaciones que el nuevo módulo deberá de cumplir. En el capítulo 5 se detallará la metodología que permite detectar las tendencias, y finalmente el capítulo 6 presentará el plan de implantación de la herramienta en la plataforma, fijado para septiembre 2020. Por último, se harán conclusiones generales y futuras líneas de trabajo para el futuro del módulo.

## Palabras clave

Vigilancia; predicción; tendencia; temáticas éticas; riesgo no financiero; análisis; red social; preense; informes financieros; sostenibilidad; medio ambiente; inversores; procesamiento natural del lenguaje; inteligencia artificial



# Resum

Datamaran és una empresa creada en 2014 a Londres, amb l'objectiu de facilitar a les grans empreses una eina innovadora que permet llegir automàticament i en temps real els documents útils per a prendre decisions estratègiques en el context mediambiental i social. El seu departament tècnic es basa quasi exclusivament a València, on treballen al voltant de 20 persones, desenvolupadors i data scientists per la majoria. La resta dels seus empleats (50 persones en total) es reparteix entre les oficines de Londres i Nova York, on s'encarreguen de les vendes comunicació, associacions i atenció al client. Dins dels clients de Datamaran es troben la BBVA, el banc Santander, Rakuten, Amazon, Philips i Morgan Stanley.

El treball s'emporta sobre un mòdul específic de la solució de Datamaran: el *continuous monitoring*. Aquest mòdul s'encarrega d'observar com van evolucionar les dades en els últims 6 mesos, perquè els clients puguem entendre millor quals són les possibles raons que expliquen perquè una temàtica donada es troba més important hui dia. S'emportarà un diàleg amb el departament business de Datamaran per a definir correctament els *insights* interessants per als clients, es definiran indicadors mesurables que permeten identificar tendències en les dades i intentar predir que temàtiques més probablement seran importants en el futur.

El document es dividirà en 7 capítols, després de la introducció, en el capítol 2 es tractarà de l'entorn del problema, en l'empresa, posant el mòdul en el seu context en l'aplicació. Després, en el capítol 3 es durà a terme una anàlisi detallada del funcionament actual del mòdul existent i en el capítol 4, gràcies al departament responsable del producte es determinaran les especificacions que el nou mòdul haurà de complir. En el capítol 5 es detallarà la metodologia que permet detectar les tendències, i finalment el capítol 6 presentarà el pla d'implantació de l'eina en la plataforma, fixat per a setembre 2020. Per últim, es faran conclusions generals i futures línies de treball per al futur del mòdul.

## Paraules clau

Vigilància; predicció; tendència; temàtiques ètiques; risc no financer; anàlisi; xarxa social; premsi; informes financers; sostenibilitat; medi ambient; inversors; processament natural de l'enguatge; intel·ligència artificial



# Summary

Datamaran is a company created in 2014 in London, with the aim of providing large companies with an innovative tool that allows them to read automatically and in real time the documents that are useful for making strategic decisions in the environmental and social context. Its technical department is based almost exclusively in Valencia, where about 20 people work, most of them developers and data scientists. The rest of its employees (50 people in total) are divided between the London and New York offices, where they are responsible for sales communication, partnerships and customer service. Datamaran's customers include BBVA, Banco Santander, Rakuten, Amazon, Philips and Morgan Stanley.

The work is carried out on a specific module of Datamaran's solution: the continuous monitoring. This module is in charge of observing how the data evolved in the last 6 months, so that customers can better understand which are the possible reasons that explain why a given topic is more important today. A dialogue with Datamaran's business department will be carried out in order to correctly define the interesting insights for the customers. Measurable indicators will be defined to identify trends in the data and to try to predict which topics are more likely to be important in the future.

The document will be divided in 7 chapters, after the introduction, in chapter 2 we will deal with the environment of the problem, in the company, putting the module in its context in the application. Then, in chapter 3 a detailed analysis of the current functioning of the existing module will be carried out and in chapter 4, thanks to the department responsible for the product, the specifications that the new module will have to meet will be determined. Chapter 5 will detail the methodology for detecting trends, and finally Chapter 6 will present the plan for implementing the tool in the platform, set for September 2020. Finally, general conclusions and future lines of work for the future of the module will be made.

## Keywords

Monitoring; prediction; trend; ethical issues; non-financial risk; analysis; social network; news articles; financial reports; sustainability; environment; investors; natural language processing; artificial intelligence



## DOCUMENTOS CONTENIDOS EN EL TFM

- Memoria
- Anejos
- Presupuesto

# Tabla de contenido

1.	Introducción.....	1
1.1.	Objeto del trabajo.....	1
1.2.	Antecedentes teóricos .....	1
1.2.1.	AMFE.....	1
1.2.2.	Scrum.....	2
1.2.3.	Ley de Zipf.....	3
1.2.4.	Python .....	5
1.3.	Estructura del documento .....	5
2.	Descripción del entorno .....	7
2.1.	Introducción.....	7
2.2.	La empresa Datamaran Erevalue SLU .....	7
2.3.	Los clientes y proveedores de la empresa.....	10
2.3.1.	Clientes .....	10
2.3.2.	Proveedores.....	11
2.4.	Productos y servicios .....	12
2.4.1.	Introducción a la plataforma .....	12
2.5.	Organización interna.....	14
2.5.1.	Organización general.....	14
2.5.2.	Organización de la oficina de Valencia .....	15
2.5.3.	Procesos de trabajo.....	16
2.6.	Conclusión .....	19
3.	Análisis de la situación actual .....	21
3.1.	Introducción.....	21
3.2.	Desarrollo de un análisis de materialidad .....	21
3.2.1.	Resultados visibles en <i>Matrix View</i> .....	21
3.2.2.	Parametrización y cálculo de las fuentes .....	27
3.2.3.	Parámetros del Benchmark.....	30
3.2.4.	Leyes y recomendaciones legales .....	34
3.2.5.	Noticias y redes sociales.....	39
3.3.	Continuous monitoring .....	46
3.3.1.	Caso de uso.....	46



3.3.2.	Arquitectura y procesos involucrados .....	49
3.4.	Problemas y oportunidades encontrados.....	49
3.5.	Conclusión .....	50
4.	Definición de los requerimientos del módulo de monitoring .....	53
4.1.	Introducción.....	53
4.2.	Actores del proyecto .....	53
4.3.	Requerimientos obligatorios.....	54
4.3.1.	Requerimientos funcionales.....	54
4.3.2.	Requerimientos no funcionales.....	58
4.4.	Requerimientos libres.....	60
4.5.	Conclusión .....	62
5.	Definición de los indicadores.....	63
5.1.	Introducción.....	63
5.2.	Metodología .....	63
5.2.1.	Diagrama de flujo .....	63
5.2.2.	Hipótesis .....	65
5.3.	Señales sociales .....	67
5.3.1.	Criterios de selección .....	67
5.3.2.	Definición de las alternativas.....	71
5.3.3.	Evaluación y selección .....	72
5.3.4.	Crítica de la selección .....	74
5.4.	Tendencia .....	75
5.4.1.	Criterios de selección .....	76
5.4.2.	Definición de las alternativas.....	78
5.4.3.	Selección .....	81
5.4.4.	Crítica de la selección .....	89
5.5.	Conclusión .....	90
6.	Implantación de la herramienta.....	91
6.1.	Introducción.....	91
6.2.	Plan de implantación.....	91
6.3.	Plan de contingencias .....	95
6.4.	Valoración económica .....	98
6.5.	El nuevo módulo.....	100
6.6.	Conclusión .....	105
7.	Conclusiones .....	107
7.1.	Conclusión de la memoria.....	107
7.2.	Lecciones aprendidas .....	107



7.3. Futuras líneas de trabajo.....	108
Bibliografía.....	109
Anejos.....	111
Anejo A – Código Python Parte 1.....	111
Anejo B – Código Python Parte 2.....	114
Anejo C – Código Python Parte 3.....	117
Anejo D – Ontología: definición de los tópicos .....	120
Presupuesto .....	121
Contenido .....	121
Consideraciones previas .....	123
Justificación de precios por capítulo .....	125
Presupuesto de ejecución material .....	129
Presupuesto por contrata.....	131

# Índice de ilustraciones

Ilustración 1 - Estructura de los Sprints en método Scrum. (Fuente: (Proyectos Ágiles)).....	3
Ilustración 2 - Distribución de las palabras en inglés. Frecuencia de uso en función del ranking - escala log, log. (Fuente: (Rodríguez Abellán, 2019)).....	4
Ilustración 3 - Población de las ciudades en función de su rango para 8 países diferentes (escala log-log). (Fuente: (Rodríguez Abellán, 2019)).....	4
Ilustración 4- Frisa cronológica de la creación de la empresa Datamaran Ltd. y filiales.(Fuente: elaboración propia).....	7
Ilustración 5 - Ubicación de las oficinas de Datamaran en el espacio de coworking WayCo (El Carmen) (Fuente: Google Maps).....	8
Ilustración 6- Sala de reuniones (Sala el Carmen) llena durante la retrospectiva semanal del viernes (Fuente: elaboración propia) .....	9
Ilustración 7 - Ubicación de las nuevas oficinas de Datamaran Erevalue SLU en Valencia. (Fuente: Google Maps).....	9
Ilustración 8 - Obras en curso en la nueva oficina de Datamaran (dirección Poeta Querol, Valencia). (Fuente: Datamaran) .....	10
Ilustración 9 – Algunos clientes de Datamaran. Los clientes y proveedores de la empresa Fuente: (Datamaran Ltd., 2020) .....	10
Ilustración 10 - Proveedores principales de Datamaran. (Fuente: elaboración propia) .....	11
Ilustración 11 - Pestañas de la plataforma Datamaran 4. (Fuente: plataforma Datamaran) .....	12
Ilustración 12 - Ejemplo de gráfico de materialidad disponible en la pestaña "Materiality" de la aplicación web Datamaran 4. (Fuente: Datamaran) .....	13
Ilustración 13 - Organigrama general de la dirección de Datamaran. (Fuente: elaboración propia).....	14
Ilustración 14 - Esquema de los sistemas de comunicación interna de Datamaran. (Fuente: elaboración propia).....	15
Ilustración 15 - Organigrama simplificado de la composición del departamento técnico de Datamaran. (Fuente: elaboración propia) .....	16
Ilustración 16 - Ejemplo de user story en JIRA. (Fuente: JIRA software). .....	17
Ilustración 17 - Tabla individual de progresión del sprint en JIRA. (Fuente: JIRA software) .....	17
Ilustración 18 - Ejemplo de gestión de versiones en BitBucket gracias al concepto de ramas. (Fuente: BitBucket-SourceTree) .....	18
Ilustración 19 - Presentación de la evolución del entorno durante el proceso de calidad para desplegar una aplicación web. (Fuente: elaboración propia). .....	19
Ilustración 20 - Ejemplo de análisis de materialidad. Verticalmente: importancia externa. Horizontalmente: Importancia interna. (Fuente: Datamaran).....	22
Ilustración 21 - Algunos valores de Amazon. (Fuente: Amazon Spain) .....	23
Ilustración 22 - Ejemplo de matriz de rangos materialidad. Columnas de izquierda a derecha: nombre de la temática, Benchmark, Regulaciones obligatorias, Recomendaciones legales, Noticias, Redes sociales (Twitter), Score externo, Score externo y Relevancia. (Fuente: Datamaran) .....	24
Ilustración 23 - Página de selección de un análisis por el usuario en el módulo de materialidad. (Fuente: Datamaran) .....	24
Ilustración 24 - Vista interna de parametrización de la matriz. (Fuente: Datamaran) .....	25
Ilustración 25 - Esquema resumen de los cálculos realizados para evaluar la importancia externa de una temática. (Fuente: elaboración propia).....	27
Ilustración 26 - Ejemplo de mapa de tópicos para la temática Supplier Management. (Fuente: Datamaran).....	28



Ilustración 27 - Pestaña Materiality/External view, dónde se entran los parámetros del enfoque para cada fuente. (Fuente: Datamaran) .....	29
Ilustración 28 - Pestaña de definición del perfil de usuario. (Fuente: Datamaran) .....	29
Ilustración 29 - Pestaña Benchmark libre para la exploración de los informes de los competidores. (Fuente: Datamaran).....	30
Ilustración 30 - Parámetros modificables por el usuario en la pestaña de benchmark. (Fuente: Datamaran) .....	31
Ilustración 31 - gráfico de nivel de mención por tópico en los informes anuales de las empresas de tamaño grande del sector agroalimentario en Europa. (Fuente: Datamaran) .....	32
Ilustración 32 - Diferentes gráficos por tópico de la pestaña Benchmark. De arriba hacia abajo y de izquierda a derecha: Posición del usuario, desglose geográfico, desglose por sector de actividad económica, desglose por compañía. (Fuente: Datamaran).....	33
Ilustración 33 - gráfico perfil del usuario en la pestaña benchmark. (Fuente: Datamaran)...	34
Ilustración 34 - Panel de parametrización de Benchmark en la pestaña de materialidad. (Fuente: Datamaran) .....	34
Ilustración 35 - Pestaña Observe: formulario de selección del enfoque de las regulaciones. (Fuente: Datamaran) .....	35
Ilustración 36 - Pestaña My Regulatory Radar. Gráficos de arriba hacia abajo, de izquierda a derecha: lista de regulaciones por tipo de relevancia, tópicos con más regulaciones, regulaciones por país, lista de textos más recientes. (Fuente: Datamaran) .....	36
Ilustración 37 - Panel de parámetros para las regulaciones (obligatorias y recomendadas) en la pestaña Materiality/External view. (Fuente: Datamaran).....	37
Ilustración 38 - Ejemplo de detalles sobre una ley publicada y referenciada en la plataforma Datamaran. (Fuente: Datamaran).....	38
Ilustración 39 - Pestaña NewsFlow, vista por compañía. Gráfico: evolución del número de artículos mencionando SANOFI o PZIFER INC entre 04/2018 y 04/2020. (Fuente: Datamaran) .....	39
Ilustración 40 - Gráfico: evolución del número de artículos mencionando SANOFI o PZIFER INC entre 04/2018 y 04/2020. (Fuente: Datamaran).....	41
Ilustración 41 - Gráfico: Evolución temporal de las menciones de empresas pertenecientes a la industria de la salud y productos farmacéuticos entre 04/2018 y 04/2020. (Fuente: Datamaran) .....	41
Ilustración 42 - Gráfico: Evolución temporal del número de artículos tratando de los tópicos: polución del aire, ensayos clínicos, mecanismos de reclamaciones. (Fuente: Datamaran) .....	42
Ilustración 43 - Panel de parámetros de la fuente de noticias de prensa en la pestaña Materiality/External view. (Fuente: Datamaran) .....	42
Ilustración 44 - Gráfico: evolución del número de tuits relacionados con los tópicos acceso a la medicina y ensayos clínicos. (Fuente: Datamaran) .....	43
Ilustración 45 - Pestaña Social, vista detalles por tópico de los tuits acerca del tópico: acceso a la medicina - 20/04/2020. (Fuente: Datamaran).....	44
Ilustración 46 - Gráfico: evolución de los tuits semanales acerca del tópico: acceso a la medicina. - 20/04/2020. (Fuente: Datamaran) .....	44
Ilustración 47 - Gráfico: hashtags más populares acerca del tópico: acceso a la medicina durante la semana del 06/04/2020. (Fuente: Datamaran).....	45
Ilustración 48 - Otro ejemplo de análisis de materialidad - vista gráfica. (Fuente: Datamaran) .....	46
Ilustración 49 - Acceso al módulo de continuous monitoring desde la vista del análisis de materialidad. (Fuente: Datamaran) .....	47
Ilustración 50 - Ejemplo de vista de continuous monitoring con solo una temática incrementando. (Fuente: Datamaran).....	47



Ilustración 51 - Lista/Tabla de Continuous monitoring que presenta las 5 temáticas que incrementaron más en rango. (Fuente: Datamaran) .....	48
Ilustración 52 - Evolución del score externo de la temática Employment practices para el análisis de la Ilustración 43. (Fuente: Datamaran) .....	48
Ilustración 53 - Esquema de comunicaciones entre las diferentes partes del software Datamaran. (Fuente: elaboración propia). .....	49
Ilustración 54 - Esquema (mockup ) de la nueva disposición de la tabla de temáticas. (Fuente: elaboración propia). .....	55
Ilustración 55 - Esquema (mockup) de la situación deseada para la parte baja de la página, el gráfico. (Fuente: elaboración propia).....	57
Ilustración 56 - Esquema de comunicaciones entre las diferentes partes del software Datamaran. (Fuente: elaboración propia) .....	59
Ilustración 57 – Plantilla PowerPoint de la responsable de customer success, para que puedan comunicar los cambios importantes. Diapositiva de evolución de las temáticas. (Fuente: Datamaran).....	61
Ilustración 58 - Diapositiva presentando las señales de importancia social atribuida a ciertas temáticas. (Fuente: Datamaran). .....	62
Ilustración 59 - Diagrama de flujo del proceso de selección de un indicador. (Fuente: elaboración propia).....	64
Ilustración 60 - Ejemplo de diagnóstico estático detectando las temáticas que probablemente ganarán en importancia en los próximos 6 meses según el método elegido. (Fuente: elaboración propia).....	74
Ilustración 61 - Gráfico representando la evolución temporal de la presencia de la temática Climate change & air quality. (Fuente: elaboración propia) .....	76
Ilustración 62 - Evolución temporal de la presencia de la temática Climate change & air quality con el límite calculado en función de los valores propios de la evolución ( $t_2 = 1$ ). (Fuente: elaboración propia).....	79
Ilustración 63 - Curva representativa de la importancia acumulada con el ranking de la temática para diferentes fuentes. (Fuente: elaboración propia) .....	80
Ilustración 64 - Gráficos representando el log(presencia) para cada temática del conjunto de datos según su ranking en el conjunto total. (Fuente: elaboración propia).....	80
Ilustración 65 - Ejemplo de hoja de cálculo conteniendo las previsiones del modelo seleccionado para las tendencias. (Fuente: elaboración propia) .....	89
Ilustración 66 - gráfico de evolución del número de páginas de la memoria. (Fuente: elaboración propia).....	93
Ilustración 67 – Cronograma del proyecto de mejora del módulo de continuous monitoring (Fuente: elaboración propia) .....	94
Ilustración 68 - Panel de vigilancia de la actividad en las lambdas. (Fuente: Amazon Web Service Console) .....	95
Ilustración 69 - Ejemplo de notificación por correo electrónico de detección de un problema con las lambdas de AWS. (Fuente: Gmail).....	95
Ilustración 70 - Esquema de proceso de los unit tests. (Fuente: elaboración propia) .....	96
Ilustración 71 - Mail enviado por la VP Brand & Business de Datamaran después del Webinar de formación de los clientes acerca del nuevo módulo . (Fuente: Datamaran).....	99
Ilustración 72 - Nueva tabla del módulo de monitoring. (Fuente: Datamaran) .....	100
Ilustración 73 - gráfico en el nuevo módulo de continuous monitoring. (Fuente: Datamaran) .....	100
Ilustración 74 - Hoja "Summary" del archivo exportable del nuevo módulo de monitoring. (Fuente: Datamaran) .....	101
Ilustración 75 - Hoja "6 months view" del archivo exportable del nuevo módulo de monitoring.	



(Fuente: Datamaran) .....	102
Ilustración 76 - Ubicación del nuevo botón para exportar el archivo Excel conteniendo el análisis de temáticas presentando señales y futuras tendencias. (Fuente: Datamaran).....	103
Ilustración 77 - Ejemplo de tabla en la hoja principal del archivo Excel exportado desde Datamaran, conteniendo la detección de señales y futuras tendencias. (Fuente: Datamaran) ..	104

# Índice de tablas

Tabla 1 - Ejemplo de línea en un AMFE. (Fuente: elaboración propia) .....	2
Tabla 2 - Diferentes niveles de riesgo en función del NPR. (Fuente: (Stamatis, 2003)).....	2
Tabla 3 - Roles importantes en el método Scrum. (Fuente: (Proyectos Ágiles)) .....	3
Tabla 4 - Top 10 de los lenguajes más representados en los repositorios GitHub en el primer cuatrimestre de 2020. (Fuente: (Madnight GitHub, 2020)).....	5
Tabla 5 - Repartición de los diferentes tipos de sectores entre las diferentes industrias (en negrita). (Fuente: Datamaran).....	40
Tabla 6 - Lista de actores del proyecto y sus roles en Datamaran. (Fuente: elaboración propia) .....	53
Tabla 7 - Requerimientos para la tabla de monitoring, presentados por el equipo de producto que se tienen que realizar obligatoriamente en la nueva versión del módulo . (Fuente: elaboración propia).....	54
Tabla 8 - Requerimientos para el gráfico de continuous monitoring. (Fuente: elaboración propia).....	56
Tabla 9- Tooltip que debe aparecer al seleccionar dos puntos distintos en el gráfico de monitoring. (Fuente: elaboración propia).....	57
Tabla 10 - Correspondencia de nombre y códigos para el requerimiento R9. (Fuente: elaboración propia).....	58
Tabla 11 - Requerimientos abiertos para el nuevo módulo de monitoring, acerca del diagnóstico automático de los análisis. (Fuente: elaboración propia).....	60
Tabla 12 - Frecuencias de agregación de nuevos elementos a la base de datos. (Fuente: Datamaran).....	65
Tabla 13 - Criterios de selección de la señal para detectar temáticas de moda actualmente. (Fuente: elaboración propia) .....	67
Tabla 14 - Ejemplo de evaluación de una alternativa. (Fuente: elaboración propia) .....	69
Tabla 15 - Escala de medida de importancia relativa de criterios. (Fuente: (Saaty, 2008))..	69
Tabla 16 - Matriz de prioridad de los criterios. En negrita están los números entrados manualmente, el resto se calcula automáticamente de forma coherente. (Fuente: elaboración propia).....	70
Tabla 17 - Estimaciones del valor propio máximo de la matriz de jerarquía. (Fuente: elaboración propia).....	70
Tabla 18 -Ratio de consistencia de una matriz aleatoria en función de su dimensión. (Fuente: (Saaty, 2008)).....	70
Tabla 19 - Resultados de la optimización de los parámetros con una búsqueda local. (Fuente: elaboración propia).....	72
Tabla 20 - Resultados obtenidos por las 3 alternativas según los criterios y su respectivo peso. (Fuente: elaboración propia) .....	73
Tabla 21 - Puntuación obtenida por dos alternativas triviales. (Fuente: elaboración propia) .....	74
Tabla 22 - Ejemplo de datos usados para evaluar el desempeño de las señales. Para cada mes se usa la presencia de la temática (%). (Fuente: Datamaran).....	76
Tabla 23 - Matriz de prioridad de los criterios. En negrita están los números entrados manualmente, el resto se calcula automáticamente de forma coherente. (Fuente: elaboración propia).....	78
Tabla 24 - Resultados de la optimización de los parámetros para cada alternativa y cada tipo de fuente. (Fuente: elaboración propia).....	81
Tabla 25 - Desempeño de soluciones triviales con la previsión de tendencias sobre	



recomendaciones legales. (Fuente: elaboración propia) .....	82
Tabla 26 - Resultados de exactitud de las diferentes alternativas. (Fuente: elaboración propia) .....	82
Tabla 27 - Resultados de las diferentes alternativas sobre el conjunto de datos de recomendaciones legales. (Fuente: elaboración propia) .....	82
Tabla 28 - Datos del análisis de desempeño de la alternativa 1 sobre la Fuente: recomendaciones legales. (Fuente: elaboración propia) .....	83
Tabla 29 - Resultados esperados si el método de clasificación no cambiaría las distribuciones. (Fuente: elaboración propia) .....	84
Tabla 30 - Resultados de las diferentes alternativas sobre el conjunto de datos de artículos de prensa. (Fuente: elaboración propia).....	84
Tabla 31 - Resultados de exactitud para las diferentes alternativas y para la Fuente: artículos de prensa. (Fuente: elaboración propia) .....	85
Tabla 32 - Resultados obtenidos con las alternativas modificadas. (Fuente: elaboración propia) .....	85
Tabla 33 - Resultados obtenidos para las diferentes alternativas seleccionadas para la Fuente: artículos de prensa. (Fuente: elaboración propia) .....	85
Tabla 34 - Recuento de temáticas según si esta actividad o no, y si incrementa o no, para el test de la alternativa 4. (Fuente: elaboración propia) .....	86
Tabla 35 - Resultados esperados en caso de validación de la hipótesis H0, y p-value obtenida con los resultados reales. (Fuente: elaboración propia) .....	86
Tabla 36 - Resultados obtenidos por 3 soluciones triviales para la fuente de datos: social media. (Fuente: elaboración propia) .....	87
Tabla 37 - Resultados de exactitud para las diferentes alternativas y para la Fuente: social media. (Fuente: elaboración propia).....	87
Tabla 38 - Resultados obtenidos por las diferentes alternativas en la Fuente: social media. (Fuente: elaboración propia) .....	87
Tabla 39 - Resultados obtenidos con la clasificación de la alternativa 3 para la Fuente: social media. (Fuente: elaboración propia).....	88
Tabla 40 - Resultados esperados si la hipótesis nula fuese verdad. (Fuente: elaboración propia) .....	88
Tabla 41 - Resumen de las diferentes soluciones seleccionadas. (Fuente: elaboración propia) .....	88
Tabla 42 - Lista de responsables y participantes al proyecto de mejora de continuous monitoring. (Fuente: elaboración propia).....	92
Tabla 43 - Recursos humanos necesarios para el proyecto .....	93
Tabla 44 - Análisis de Modos de Fallos y Efectos del proyecto de mejora. (Fuente: elaboración propia).....	97
Tabla 45 - Diferentes escenarios para el retorno sobre inversión del proyecto. (Fuente: elaboración propia).....	98





# 1. Introducción

## 1.1. Objeto del trabajo

En el presente trabajo se realizará un análisis del funcionamiento de Datamaran, una plataforma de gestión de riesgos no financieros, y se implantará una mejora del módulo de vigilancia de aquellos riesgos. Para redactar su informe anual financiero, una empresa multinacional tiene que tomar en cuenta las temáticas medio ambientales y sociales más importantes para sus inversores, y actuar conforme a la importancia de esas temáticas. Por tanto, Datamaran provee una manera de monitorear esas temáticas en su módulo llamado *continuous monitoring*, pero este módulo presenta varios defectos y una falta de perspectiva y propuesta de valor. Los clientes están interesados en predicciones sobre lo que serán los riesgos futuros que tendrán que gestionar a tiempo y por tanto quieren un método que les permita focalizar su energía sobre lo que realmente importa.

Se ha seleccionado la solución de Datamaran porque actualmente son los clientes exclusivos de mi microempresa de desarrollo web, data science y *consulting* en tecnologías. A principios de 2020 era becario en Datamaran antes de pasar a ser un *contractor* independiente de la empresa.

El diseño y la implantación del nuevo módulo de *monitoring* ha sido una de mis misiones en Datamaran, y al ser una empresa de tamaño relativamente pequeño (20 personas en el departamento técnico) me dejaron proponer mis ideas e implantarlas apoyándome con todos los recursos necesarios.

## 1.2. Antecedentes teóricos

### 1.2.1. AMFE

El AMFE o Análisis de Modos de Fallos y sus Efectos es una herramienta usada en teoría del mantenimiento industrial para priorizar la gestión de los riesgos en la ejecución de un proceso. Es un método esencial para la calidad del producto, aprendiendo de los fallos anteriores para evitar que se repitan. La idea general es establecer una lista de etapas críticas y de las posibles causas que puede pasar diferente de lo previsto. Cada uno de estos fallos potenciales se van a priorizar usando 3 dimensiones:

- Gravedad: se mide la intensidad del efecto de aquel fallo sobre el proceso/la organización
- Ocurrencia: se estima la frecuencia con la cual el fallo puede ocurrir
- Detección: se mide la probabilidad que los sistemas de vigilancia implantados detectan el problema

En inglés serán *Ocurrence*, *Severity*, *Detection* (OSD) según el libro de aplicación de la teoría del AMFE (Stamatis, 2003).

Estos 3 criterios se evaluarán con una escala de 1 a 10, de menor a mayor peligro (1 en detección quiere decir que se detecta siempre el problema y 10 que nunca se detecta). Las informaciones se ponen en una tabla como la de la Tabla 1, presentando el modo de fallo de una lámpara en una nave industrial.

Tabla 1 - Ejemplo de línea en un AMFE. (Fuente: elaboración propia)

Descripción de la fase	Modo/s potencial/es de fallo	Efecto/s potencial/es del fallo	Gravedad	Causa(s) potencial(es) del fallo(s)	Ocurrencia	Verificación(es) y/o control(es) actual(es)	Detección n	NPR	Acción(es) recomendada(s)	Área(s) persona(s) responsable(s) y fecha de realización
Encender la luz	No se enciende	No se ve nada en la sala	8	Filamento quemado	4	No hay	3	96	Cambiar el bombillo	Responsable del estado de la nave

Es importante listar los modos de control y prevención de los fallos para darse cuenta de lo que se está haciendo para evitar o controlar los fallos, e identificar el área encargada de gestionar el riesgo.

Cada uno de los tres criterios (gravedad, ocurrencia y detección) se miden o se evalúan antes de servir para calcular el NPR:

$$NPR = Gravedad \cdot Ocurrencia \cdot Detección \quad (1)$$

Los riesgos de fallo se priorizan usando la escala en la Tabla 2 a continuación:

Tabla 2 - Diferentes niveles de riesgo en función del NPR. (Fuente: (Stamatis, 2003))

NPR	Nivel de riesgo
500 – 1000	Alto riesgo de fallo
125 – 499	Riesgo de fallo medio
1 – 124	Riesgo de fallo bajo
0	No existe riesgo

### 1.2.2. Scrum

El método Scrum es un método famoso en el sector informático más específicamente el desarrollo de software y aplicaciones. Se presenta como un método para trabajar en equipo de forma ágil (Proyectos Ágiles), para ello se dividen los proyectos en diferentes entregas parciales del producto final. Se organizan lo que se llaman Sprints de una o dos semanas durante las cuales cada miembro del equipo efectúa sus tareas y al final de la iteración se debe de tener una parte del producto funcionando correctamente. El esquema general del método se presenta con la Ilustración 1, cada iteración empieza con una planificación, luego todos los días se comenta lo que se está haciendo (por si alguien está bloqueado) y por fin se enseña el resultado de la iteración al cliente/*Product owner*.

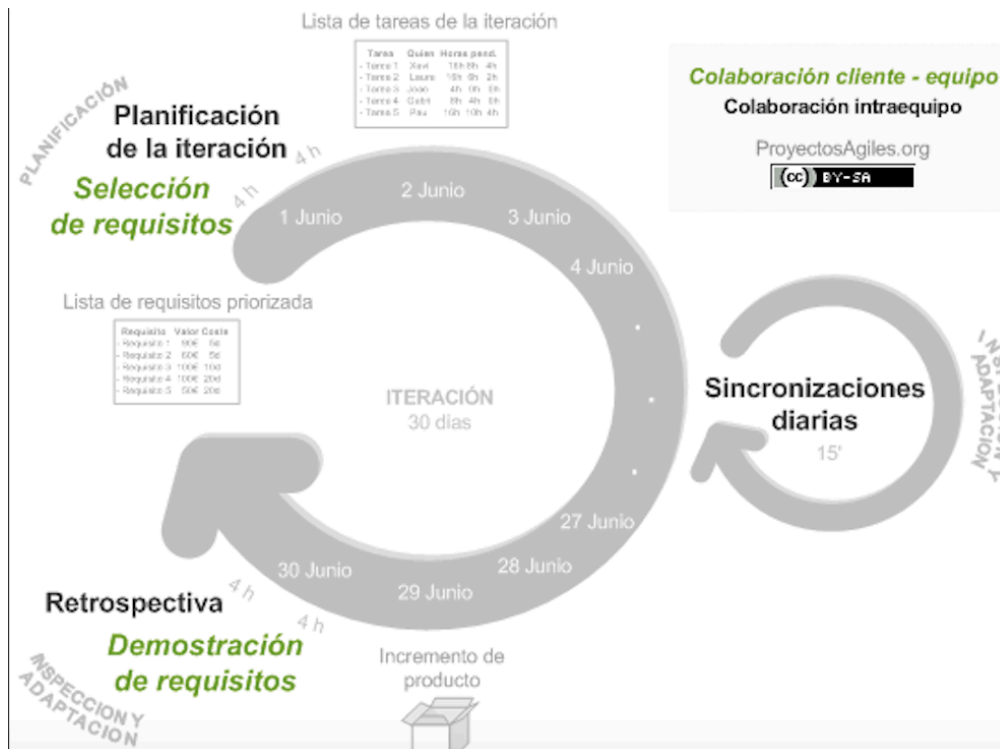


Ilustración 1 - Estructura de los Sprints en método Scrum. (Fuente: (Proyectos Ágiles))

Tabla 3 - Roles importantes en el método Scrum. (Fuente: (Proyectos Ágiles))

Roles importantes	Objetivo
<i>Scrum master</i>	Facilitar el trabajo de los miembros del equipo (eliminar obstáculos y proteger de interrupciones externas)
<i>Product Owner</i>	Prioriza los objetivos según el valor que le aportan

El método Scrum fue creado por Jeff Sutherland y Ken Schwaber, y esta documentado primero en *The Scrum Guide*, (Schwaber & Sutherland, 2017) y mas tarde Sutherland escribió *A Scrum Book* (Sutherland & Cople, 2019). Estos libros forman la base de definición de lo que es el *Scrum* sus valores y metodología.

### 1.2.3. Ley de Zipf

La ley de Zipf viene de una observación interesante que se hizo sobre los lenguajes: en cualquier idioma parece que haya una relación entre el ranking de las palabras más usadas y la frecuencia de uso de aquellas. En inglés, la palabra más empleada es *the* con más del 6% de frecuencia de uso, la segunda es *a* con aproximadamente un 3% del total de palabras usadas en inglés (Newman, 2006). El patrón que aparece en general es que la segunda palabra más usada aparece la mitad de las veces que la primera, la tercera un tercio, la cuarta un cuarto ...

De forma general aparece que hay un relación simple entre el *ranking* de la palabra y su frecuencia de uso, como lo muestra el gráfico de la. Ilustración 2.



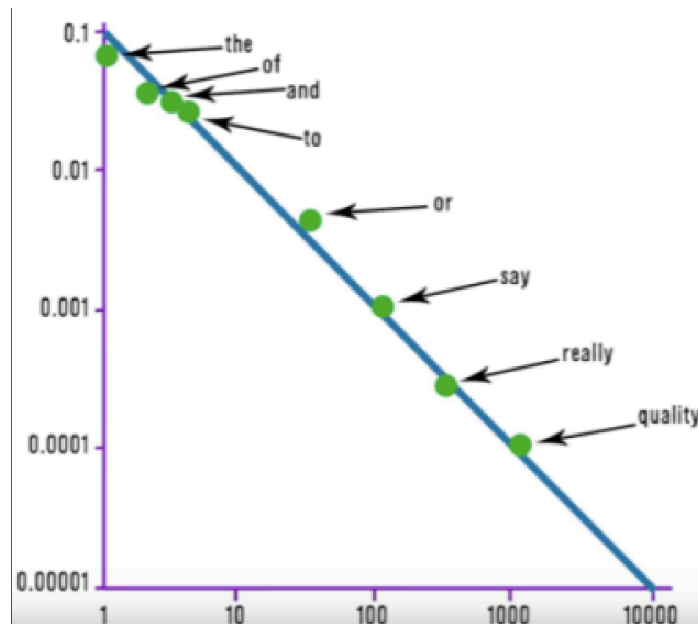


Ilustración 2 - Distribución de las palabras en inglés. Frecuencia de uso en función del ranking - escala log, log. (Fuente: (Rodríguez Abellán, 2019))

Lo más sorprendente con esta observación quizá es que se aplica a cualquier idioma (incluso los que dicen no naturales como el esperanto) y a otras áreas (como el tamaño de ciudades en la Ilustración 3), pero que no sea descubierto con certeza la causa de ello. La relación entre frecuencia de uso y ranking es una de las relaciones estudiadas en el artículo científico de la University of Michigan que estudia la detección de fenómenos que siguen una distribución parecida a la de Zipf (Newman, 2006).

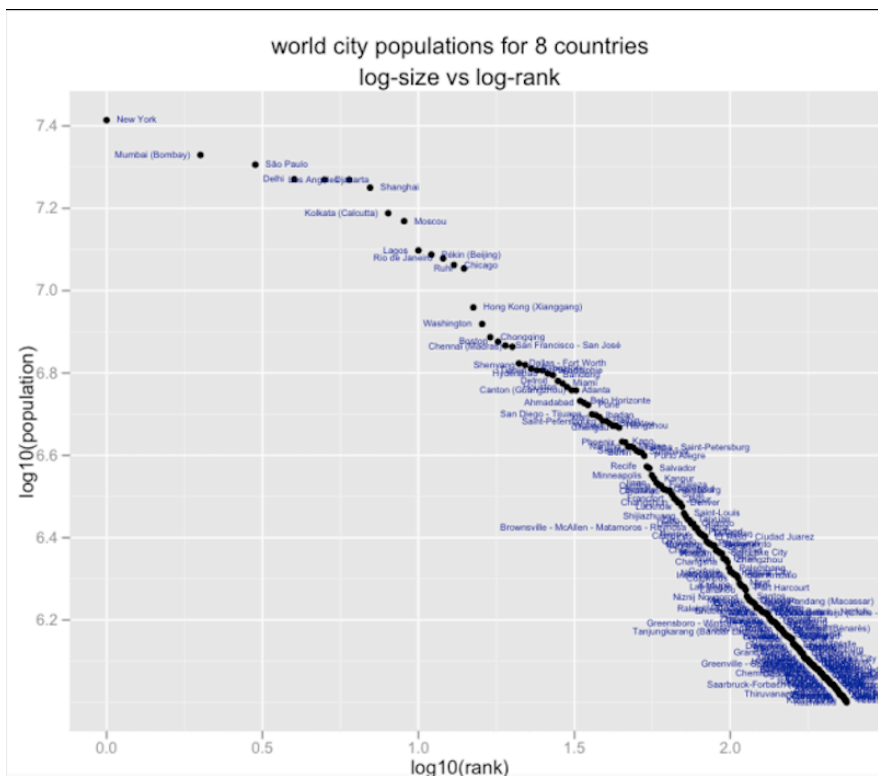


Ilustración 3 - Población de las ciudades en función de su rango para 8 países diferentes (escala log-log). (Fuente: (Rodríguez Abellán, 2019))

### 1.2.4. Python

Python es uno de los lenguajes de programación más demandados por diferentes razones (Towards Data Science, 2020):

- La sintaxis del lenguaje es elegante, simple y potente: es muy fácil escribir un código Python sin tener que preocuparse mucho por la sintaxis y se lee fácilmente.
- El conjunto de bibliotecas para ciencia de datos, o matemáticas (Numpy, Pandas, Scikit-Learn) facilita muchas operaciones a la hora de manipular y presentar datos.
- Su comunidad es una de las más activas hoy en día

Por una parte, es el lenguaje de excelencia para hacer análisis de datos, modelos de inteligencia artificial. Por otra parte, se facilita cada vez más la creación de aplicaciones autónomas enteramente en Python, gracias a los *frameworks* como Django y los proveedores de *Cloud computing* (como Amazon Web Service, Google Cloud y Microsoft Azure). Los últimos proveen servicios como las *Lambdas* (Amazon) que permiten ejecutar código del lado del servidor a bajo coste y se pueden escribir en Python.

Tabla 4 - Top 10 de los lenguajes más representados en los repositorios GitHub en el primer cuatrimestre de 2020. (Fuente: (Madnight GitHub, 2020))

# Ranking	Programming Language	Percentage (Change)	Trend
1	JavaScript	18.703% (-1.406%)	
2	Python	16.238% (-1.654%)	
3	Java	10.938% (+0.538%)	
4	Go	9.005% (+0.978%)	
5	C++	7.423% (+0.040%)	
6	Ruby	6.812% (+0.342%)	
7	TypeScript	6.769% (+1.522%)	
8	PHP	5.127% (-0.458%)	
9	C#	3.835% (+0.141%)	
10	C	3.181% (-0.203%)	

La Tabla 4 muestra lo que se acaba de explicar, Python queda segundo en GitHub detrás del versátil JavaScript (ayudado por el *framework* Node.js) que permite una ejecución de las más rápidas gracias al motor V8 de Google.

## 1.3. Estructura del documento

El trabajo se lleva a cabo sobre un módulo específico de la solución de Datamaran: el *continuous monitoring*. Este módulo se encarga de observar cómo evolucionaron los datos en los últimos 6 meses, para que los clientes puedan entender mejor cuáles son las posibles razones que explican por qué una temática dada se encuentra más importante hoy en día. Se llevará a cabo un diálogo con el departamento *business* de Datamaran para definir correctamente los *insights* interesantes para los clientes, se definirán indicadores medibles que permitan identificar tendencias en los datos y una metodología claramente explicada y justificada para la determinación de las



posibles causas de variaciones.

El documento se dividirá en 10 capítulos, después de la introducción en el capítulo 2 se tratará del entorno del problema, en la empresa, poniendo el módulo en su contexto en la aplicación. En el capítulo 3 se analizarán los antecedentes de este tipo de problema y las herramientas matemáticas, informáticas y de gestión a disposición para realizar las funcionalidades necesarias. Luego, en el capítulo 4 se llevará a cabo un análisis detallado del funcionamiento actual del módulo existente y en el capítulo 5, gracias al departamento responsable del producto se determinarán las especificaciones que el nuevo módulo deberá de cumplir. En el capítulo 6 se detalla la metodología que permite definir los indicadores, en el capítulo 7 se diseñarán todas las funcionalidades del módulo y los diferentes casos de uso para el testeo en fase final, y finalmente el capítulo 8 presentará el plan de implantación de la herramienta en la plataforma, fijado para septiembre 2020.

## 2. Descripción del entorno

### 2.1. Introducción

Datamaran Ltd. es una empresa del sector Software-as-a-Service (SaaS)<sup>1</sup> para la gestión de riesgos no financieros, creada en 2014 en Nueva York. Dentro de los clientes de Datamaran se encuentran la BBVA, el banco Santander, Rakuten, Amazon, Philips y Morgan Stanley.

El objetivo de este capítulo es presentar la empresa en su funcionamiento general para entender mejor su cadena de valor, así como el funcionamiento del software en sí, para mejor situar el trabajo en su entorno.

En este capítulo se describe la empresa Datamaran, que constituye el objeto del trabajo, más precisamente su organización interna, sus clientes y el funcionamiento global del Software que provee a sus clientes a través de su plataforma.

### 2.2. La empresa Datamaran Erevalue SLU

Datamaran Erevalue SLU es una sociedad hija de la sede central de Londres: Datamaran Ltd. Aquella fue creada en 2014 por Marjella Alma (CEO), Jean-Philippe Lecourt (CFO) y Jerome Basdevant (CTO) en Inglaterra, ex directora de GRI<sup>2</sup> y consultor en tecnología y análisis respectivamente.

Su departamento técnico se basa casi exclusivamente en Valencia, donde trabajan alrededor de 20 personas, desarrolladores y data scientists por la mayoría. El resto de sus empleados (50 personas en total) se reparte entre las oficinas de Londres (Datamaran Ltd.) y Nueva York (Datamaran Inc.), dónde se encargan de las ventas, comunicación, *partnerships*, y atención al cliente. Sus operaciones se desarrollan por la mayoría en Estados Unidos y Londres ya que decidió analizar primero el idioma más representado en el *business* de las empresas multinacionales, pero después de la implantación de las oficinas de su equipo técnico en Valencia, la empresa empezó un desarrollo fuerte en el mercado español (se inició un proyecto de análisis de documentos en castellano).

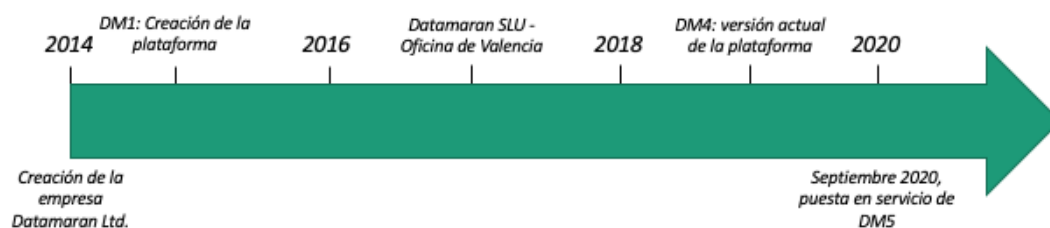


Ilustración 4- Frisa cronológica de la creación de la empresa Datamaran Ltd. y filiales. (Fuente: elaboración propia).

<sup>1</sup>Un sistema SaaS o Software as a Service, es un modelo de distribución de software donde la empresa no se encarga del almacenamiento de los datos – no posee los servidores, sino que centraliza sus datos en servidores gestionados por un proveedor externo (Amazon Web Service en el caso de Datamaran)

<sup>2</sup> Global Reporting Initiative (GRI) es una institución que crea un estándar de medidas para evaluar el desempeño de las empresas a nivel económico, ambiental y social.

La Ilustración 4 representa los momentos importantes de la vida de la empresa. Cada año desde 2015 se celebra en septiembre el lanzamiento de una nueva versión de la plataforma, en 2020 se celebrará su quinta versión.

La empresa se creó después de constatar que hoy en día la ética en las inversiones está ocupando más importancia que nunca, como lo corrobora el artículo de Bankia *“La inversión socialmente sostenible o responsable es una revolución que tiene la magnitud de la revolución industrial, pero que se mueve con la velocidad de la revolución digital y lo cierto es que es uno de los segmentos de más rápido crecimiento dentro de la industria de la gestión de activo.”* (Bankia, 2019).

Las calificaciones de inversiones éticamente validadas permiten orientar los inversores hacia empresas que les corresponden más desde el punto de vista ético, pero siempre existe el problema de la medición de la ética y de los valores de una empresa. La calificación de inversión socialmente responsable (ISR), por ejemplo, atribuye una nota ética basada en 3 pilares: ambiental, social y de gobernanza, o ESG para Environmental, Social and Governance en inglés (Label ISR, s.f.). Aquella nota sirve mucho para la toma de decisión de los inversores y fondos de inversión a la hora de elegir las empresas que financiar. Para alcanzar este objetivo, es necesario leer, anotar y sintetizar una cantidad tremenda de documentos para obtener una visión objetiva de las temáticas importantes en el entorno del cliente.

A base de esta observación, la propuesta de valor de Datamaran es de proveer a las empresas una manera de vigilar en tiempo real su posicionamiento sobre las temáticas ESG con respecto a

su entorno (competidores, leyes y opinión pública). En 2015, los fundadores comenzaron a diseñar una solución de Inteligencia Artificial (Artificial Intelligence <sup>3</sup> – AI) que proporciona información rápida eficiente y poderosa acerca de los riesgos que sus clientes necesitan identificar, evaluar y gestionar. El objetivo es permitir una gestión factual y cómoda de las temáticas no financieras relacionadas con las temáticas ESG. El software explota sobre todo las tecnologías de inteligencia artificial orientada hacia el NLP (Natural Language Processing) que permite que los algoritmos sean capaces de tener un nivel semántica de comprensión de los textos y reconocer conceptos relacionados con temáticas y empresas dentro de todo el texto que se le da para leer. Gracias a esta tecnología, se puede analizar de manera cuantitativa cuáles son las temáticas más actuales y referenciadas en un contexto dado.

Con lo cual, la misión de Datamaran es ayudar a las grandes empresas a gestionar de la mejor

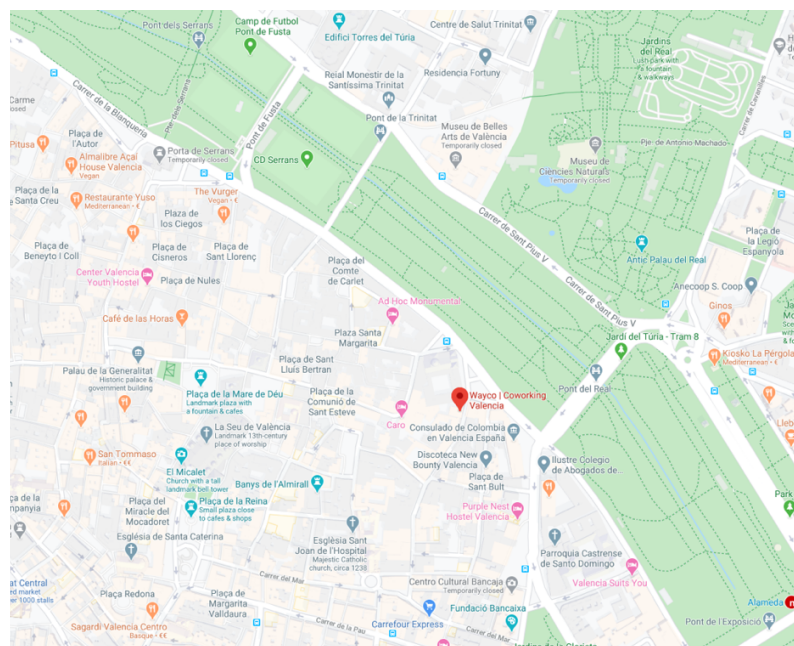


Ilustración 5 - Ubicación de las oficinas de Datamaran en el espacio de coworking WayCo (El Carmen) (Fuente: Google Maps)

<sup>3</sup> Con inteligencia artificial se refiere más precisamente a los algoritmos basados en redes neuronales de tipo GAN (*Generative Adversarial Neural networks*)



manera posible sus riesgos ESG.



Ilustración 6- Sala de reuniones (Sala el Carmen) llena durante la retrospectiva semanal del viernes (Fuente: elaboración propia)

2020 es un momento importante para la oficina española de Datamaran, ya que firma su independencia alquilando por primera vez sus oficinas reservadas únicamente a la empresa. Antes del confinamiento debido a la pandemia de Coronavirus, la empresa pagaba un acceso a un espacio ubicación de las oficinas de Datamaran cuando estaba instalada en WayCo, en la dirección calle del Gobernador Vell,27. Las razones principales de la mudanza fueron por una parte que WayCo planeaba cerrar sus oficinas del Carmen definitivamente el 30 de Marzo 2020 y por otra parte el crecimiento de los efectivos de Datamaran (de 5 a 23 empleados en Valencia entre Enero 2019 y Enero 2020) sugerían una necesidad de tener oficinas propias con más espacio, cómo se puede observar en la Ilustración 6. Debido a la crisis sanitaria de primavera 2020, la mudanza se aplazó y Datamaran entró en un periodo de teletrabajo a partir del 11 de marzo. La alta digitalización de todos los procesos internos y la división geográfica del equipo ya existente (algunos desarrolladores trabajaban desde Londres, Alemania o incluso la India) facilitaron la adaptación a un trabajo enteramente remoto.

Las futuras oficinas de Datamaran en Valencia se encuentran Calle del Poeta Querol,9 Ilustración 7, serán privadas y reservadas a Datamaran solamente, en disposición de *open-space* para facilitar la comunicación entre los diferentes miembros de los equipos de desarrollo, cómo se puede observar en la Ilustración 8.

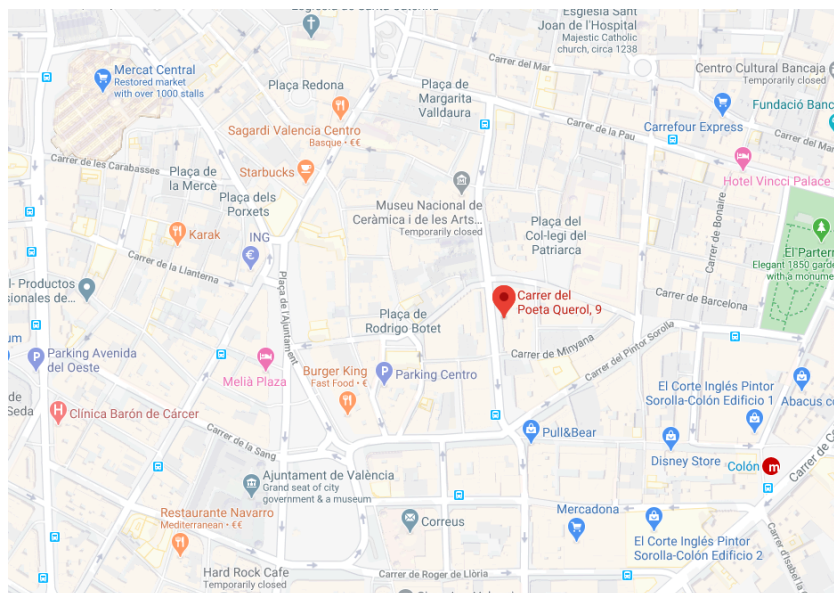


Ilustración 7 - Ubicación de las nuevas oficinas de Datamaran Evalue SLU en Valencia. (Fuente: Google Maps).



Ilustración 8 - Obras en curso en la nueva oficina de Datamaran (dirección Poeta Querol, Valencia). (Fuente: Datamaran)

## 2.3. Los clientes y proveedores de la empresa

### 2.3.1. Clientes



Ilustración 9 – Algunos clientes de Datamaran. Los clientes y proveedores de la empresa  
Fuente: (Datamaran Ltd., 2020)

A continuación, se presenta una lista de algunos clientes más importantes de Datamaran, se puede observar que todos son lo que se llama *blue-chip companies*. En 2020 Datamaran tiene

alrededor de 50 – 60 clientes que pagan el abono anual de 35.000 \$ para tener acceso a su cuenta en la plataforma y la posibilidad de realizar análisis de materialidad.

Las necesidades de los clientes no son las mismas, por tanto, la plataforma trata de ser lo más flexible posible frente a la variabilidad de la demanda, para que todos los clientes encuentren lo que les hace falta, sin que el software sea complicado de utilizar. En este proyecto, los clientes como Amazon, Philips, AT & T son las empresas que más usan el módulo de *continuous monitoring*.

### 2.3.2. Proveedores



*Ilustración 10 - Proveedores principales de Datamaran. (Fuente: elaboración propia)*

Como he explicado antes, realizar un análisis de materialidad requiere muchos datos textuales para poder realizar una clasificación de la presencia de las temáticas. Para cada fuente el proceso es diferente, muchos son automáticos y emplean pipelines (proveedores de datos). En el caso de “Benchmark” y “Newsflow”, se usa el servicio de *LexisNexis* para recolectar informes anuales corporativos y artículos de prensa. Para Twitter se usa el API de Twitter que permite pedir directamente a la red social la recolección de sus datos y tratarlos internamente después. Para las leyes y regulaciones (tanto viniendo de las autoridades nacionales, como internacionales o también ONG que escriben recomendaciones), Datamaran tiene un equipo propio que se encarga de recolectar manualmente los textos de ley que se publican para almacenarlos antes de ser leídos por el algoritmo. Para los servicios de pipeline, se abona mensualmente una cuota correspondiente al servicio.

Una vez los datos recibidos se tienen que almacenar en bases de datos, Datamaran usa principalmente *MongoDB* (Ilustración 10) como servicio de gestión de bases de datos y también *Amazon Redshift* para los datos de Twitter.

Finalmente, como empresa de SaaS, Datamaran emplea un servicio de nube para gestionar y alojar espacio para su plataforma, más precisamente el *cloud* Amazon Web Service (AWS - Ilustración 10) constituye un proveedor principal para Datamaran para todo lo que toca al despliegue de la plataforma. Por ejemplo, cuando un usuario quiere exportar los resultados de su análisis a una hoja Excel, AWS deja la posibilidad de almacenar ficheros en línea para poder ser descargado por cualquier usuario gracias a su servicio S3.



## 2.4. Productos y servicios

### 2.4.1. Introducción a la plataforma

Por una parte, el servicio ofrecido es muy innovador: se trata de usar tecnología bastante avanzada – el Natural Language Processing usando redes neuronales – para producir una herramienta perfeccionada de análisis de lo que siempre ha necesitado muchas mentes humanas para sintetizar toda la información disponible y separar lo relevante de lo irrelevante a la hora de observar las dinámicas no financieras del entorno de una empresa.

La gestión de riesgos no financieros no solo es difícil por el alto nivel de incertidumbre en el resultado de una decisión tomada sino también en la recolección de los datos importantes que llevan a tomar una decisión racional para la estrategia de gestión de aquellos riesgos. Internet da acceso “fácil” a la información: informes financieros o de sostenibilidad de las empresas, artículos de prensa, normativa reguladora de las políticas ESG de las empresas son accesibles de manera gratuita o casi gratuita en línea. No obstante, la recolección continua de estos datos - estar informado en tiempo real de cuando sale un artículo tratando de un tema dado o de una empresa en particular - y leerlos todos para estar seguros de que se trata bien del tema que nos interesa se revela ser una tarea tremendamente larga y costosa como para contratar una consultoría que se encargará de hacer la síntesis de datos. Frente a esta necesidad de mercado Datamaran se propone proveer un servicio que responderá de manera adecuada a la pregunta de las empresas multinacionales: “Cuáles son las temáticas ESG más importante para mi empresa dentro de su sector?”. Su solución se presenta como un software accesible únicamente para sus clientes desde su plataforma web <https://pro.datamaran.com/>.



Ilustración 11 - Pestañas de la plataforma Datamaran 4. (Fuente: plataforma Datamaran)

La Ilustración 11 muestra que la plataforma contiene 8 pestañas, dejando a parte “Dashboard”, “Profile” y “Search” que tienen un funcionamiento independiente, solo quedan las fuentes de información : “Benchmark” (competidores), “Observe” (leyes), “Social” (Twitter) y “Newsflow” (prensa) y la pestaña de materialidad o “Materiality”. Las pestañas tipo “fuente” permiten ajustar los parámetros del análisis para enfocar a un entorno específico (industria especial, países donde el cliente está presente, sus proveedores), la pestaña de materialidad toma en cuenta el ajuste hecho en las pestañas de fuentes para generar una síntesis de las temáticas.

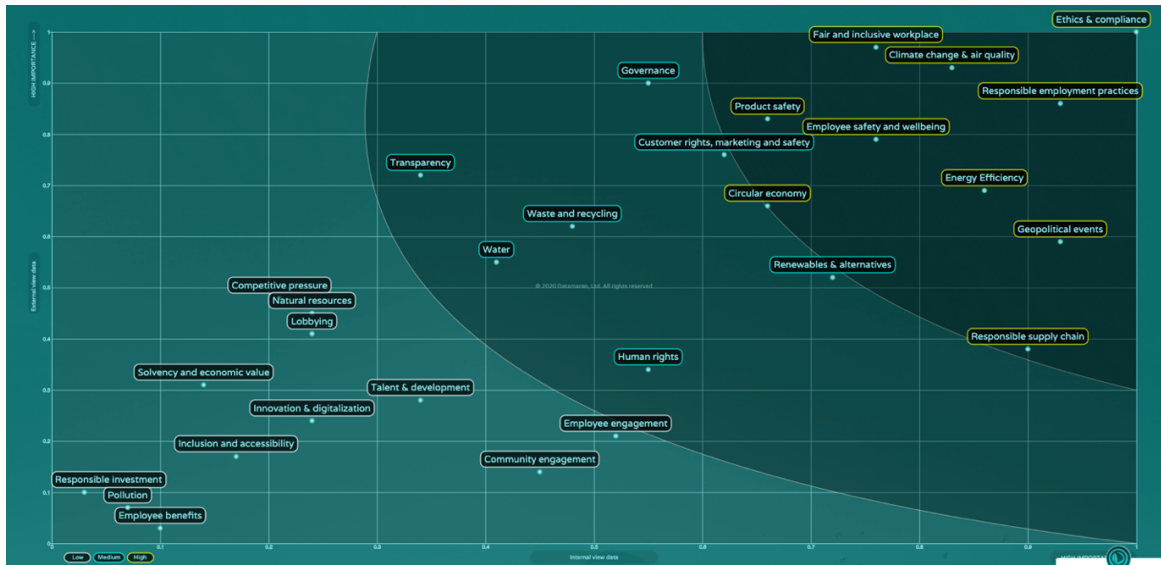


Ilustración 12 - Ejemplo de gráfico de materialidad disponible en la pestaña "Materiality" de la aplicación web Datamaran 4. (Fuente: Datamaran)

El objetivo final de la pestaña “Materiality” es generar el gráfico presentado en Ilustración 12. Este gráfico pone en perspectiva por una parte (eje horizontal), la importancia relativa de las temáticas desde el punto de vista de la empresa. El cliente tiene la posibilidad de clasificar – según qué importancia ellos le están atribuyendo – cada temática a lo largo de este eje. Gracias a un método de clasificación y cálculo de *score* se determina una clasificación “externa” de las temáticas. De arriba hacia abajo aparecen las temáticas más importantes hacia las menos importantes en el entorno definido en la plataforma. Este gráfico se puede interpretar como un diagrama de riesgo de las temáticas, idealmente, todas las temáticas se deberían de encontrar en la diagonal, significando que la política interna de la empresa acerca de la temática está en adecuación con su importancia externa. Las zonas oscuras definen zonas de relevancia, arriba a la izquierda se sitúan las temáticas más importantes de controlar, pero que en general ya están controladas, luego las de relevancia media, pueden significar un riesgo (ej.: *Climate change* aquí) por ser más importante externamente que lo son para la política de la empresa. Finalmente están las temáticas menos relevantes, que o bien no están en el foco de la empresa, o bien no son importantes externamente. Se puede explicar por dos razones, o bien la empresa no se siente impactada por la temática y estima que su riesgo es mínimo, o bien no se ha dado cuenta de que en su entorno esta temática era bastante importante y debería actuar rápidamente para corregir su posición. Estas últimas serán las más importante a la hora de vigilar los comportamientos de las temáticas a lo largo del tiempo.

### ¿Cómo determinar si un texto contiene una empresa y una temática?

Equipo de data scientists y expertos lingüistas que definen una ontología (conjunto de conceptos y palabras claves que definen un tópico), y gracias a SpaCy: librería Python de *Natural Language Processing (NLP)* se puede acceder a un conjunto de herramientas (funciones, métodos) para anotar un texto cuando se detectan los conceptos. Gracias a SpaCy se puede entrenar un modelo de Deep Learning sobre largos corpus de texto para reconocer e interpretar el texto que se le da como input.

## 2.5. Organización interna

### 2.5.1. Organización general

La organización general de la empresa se puede sintetizar como en la Ilustración 10, los 3 fundadores, sus puestos y los departamentos bajo su responsabilidad. La parte conteniendo todos los *stakeholders* del proyecto será la parte dirigida por el CTO, Jérôme Basdevant (también mi *manager* y tutor de prácticas). Más precisamente, el interesado y cliente internos del proyecto será el departamento de producto dirigido por Ian van der Vlugt ya que ellos se encargan de la buena adecuación entre expectativas de los clientes y diseño de la plataforma.

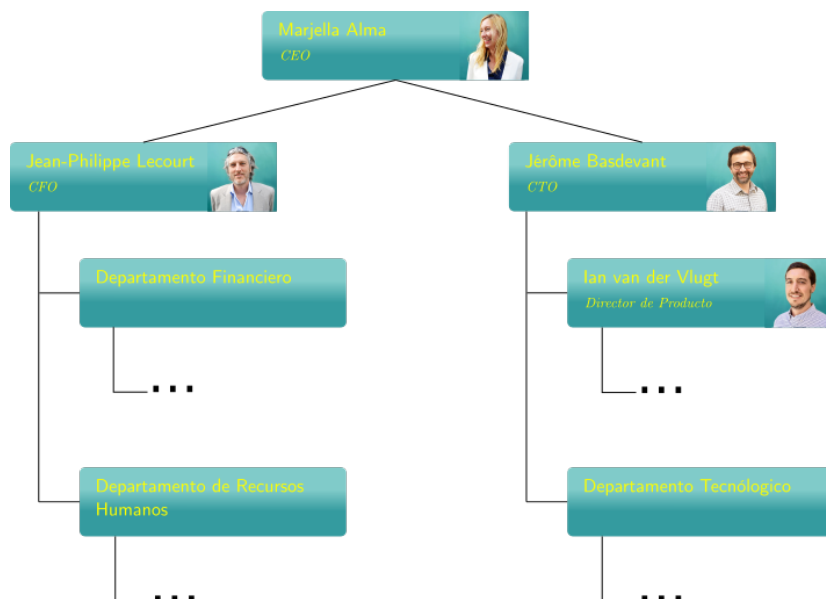


Ilustración 13 - Organigramma general de la dirección de Datamaran. (Fuente: elaboración propia)

Acerca de la comunicación interna de la empresa, hay que diferenciar los departamentos en dos categorías principales:

- Los responsables de la parte *business* (producto, ventas y marketing)
- El equipo responsable de la realización tecnológica (departamento técnico)



Ilustración 14 - Esquema de los sistemas de comunicación interna de Datamaran. (Fuente: elaboración propia).

De manera general, la mayoría de las comunicaciones pasan por Slack, cada equipo tiene su canal privado que sirve sobre todo para resolver dudas acerca de un tema especial, pedir autorizaciones a los administradores de herramientas, o en casos de teletrabajo, mantener el equipo informado de los avances, problemas encontrados el mismo día. También existe un canal interconectando los departamentos orientados *business* y el equipo tecnológico, llamado *#asktech* y permite resolver dudas, informar de la presencia de bugs en la plataforma. Para el resto de las comunicaciones interdepartamentales se usará con preferencia correos (Gmail) o Visio conferencias (Google Hangouts) como enseñado en la Ilustración 14. Las explicaciones de la comunicación interna al equipo tecnológico se detallan en el apartado 2.5.2 más adelante.

### 2.5.2. Organización de la oficina de Valencia

En la oficina de Valencia trabajan la mayoría del equipo tecnológico (2 desarrolladores trabajan desde Londres), el espacio se presenta como un *open-space* y 2 salas de reunión. El equipo tecnológico se divide en dos grupos de trabajadores principales:

- La parte *front-end* y *web design*: se encargan de la experiencia del usuario en la plataforma desde los aspectos visuales (colores y diseños) hasta la implementación de una buena lógica de recolección de los datos que el usuario espera ver, de la manera más gráfica posible.
- La parte *back-end* y *data science*, la mayoría alternan entre los dos roles.

Un desarrollador *back-end* se encarga, con el código que escribe, de proveer al *front-end* los datos correctos, por ello, los buscan llamando a los servicios Datamaran ya existentes (ubicados en el *cloud* Amazon Web Service – AWS) o a las bases de datos de la empresa (usando los sistemas *MongoDB* y *Redshift*).

Los *data scientists* se encargan de misiones de investigación interna: analizar y observar los datos almacenados en las bases de datos de la empresa, para intentar entender cómo evolucionan, qué sentido dar a los datos, y posiblemente rediseñar la lógica del *back-end*. Los *NLP data scientists* se encargan más específicamente de la interpretación de las palabras leídas por los algoritmos, de la definición de la ontología<sup>4</sup>

<sup>4</sup> La ontología se refiere a la definición de los temas identificados en un texto dado, un tema se declara presente si uno de los conceptos definidos por la ontología como haciendo referencia a este tema está mencionado en aquel texto.

así como otras tareas relacionadas con la mejora de la comprensión semántica que el algoritmo tiene de una palabra, intentando que sea más humana posible.

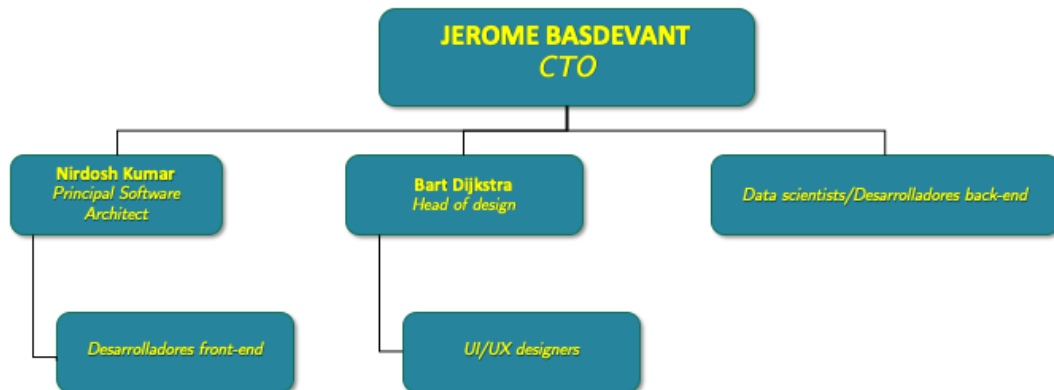


Ilustración 15 - Organigrama simplificado de la composición del departamento técnico de Datamaran. (Fuente: elaboración propia)

El organigrama de la Ilustración 15 presenta de manera sintética la organización de los equipos tecnológicos, en la realidad, los roles de cada empleado no son fijos, ya que muchos tienen varias competencias y pueden cambiar de rol si el proyecto en curso lo requiere.

### 2.5.3. Procesos de trabajo

#### Organización según sprints de desarrollo – JIRA

Como en la mayoría de las empresas de desarrollo de softwares/webs o servicios informáticos en general, el método de gestión de proyecto más usado es el método ágil. Su aplicación consiste sobre todo en dividir el producto en *features*, o funcionalidades deseadas en la plataforma. Eso significa que el desglose del proyecto en tareas se hace desde el punto de vista del usuario, y se escribe lo que se llaman *user stories*, y traducen un caso de uso de un usuario en la plataforma. Todas las *user stories* no se valen, como en el caso de Datamaran la plataforma se puede dividir en módulos, existen *epic stories*, que corresponden a bloques de *user stories* realizando en conjunto una funcionalidad más compleja, como es el caso para el módulo de *continuous monitoring*.

Cada lunes por la mañana se reúnen los equipos de *front-end* y diseño, por una parte, y de *back-end/data scientists* por otra, de cara a hacer el *sprint<sup>5</sup> planning*. La reunión de *sprint planning*, permite:

- Revisar el *sprint* anterior y comentar los diferentes problemas encontrados
- Atribuir las tareas a todos para la semana
- Añadir posibles nuevas tareas necesarias en el *backlog<sup>6</sup>*

Para realizar el seguimiento de un *sprint* dado en la empresa, se usa el software en línea de gestión de proyectos ágiles JIRA, dónde se definen las *user/epic stories* las tareas, los bugs y se atribuyen a los diferentes miembros del equipo. Un elemento de tarea en JIRA debe de tener un número de *story points* que corresponde a la dificultad de la tarea en sí (se cuantifican siempre como potencias enteras de 2).

<sup>5</sup> Un *sprint* corresponde a una semana completa de trabajo durante la cual se realizan las tareas atribuidas a cada trabajador.

<sup>6</sup> El *backlog* corresponde a la “pila” de tareas que ya han sido definidas pero que todavía no están programadas

Ilustración 16 - Ejemplo de user story en JIRA. (Fuente: JIRA software).

Ilustración 17 - Tabla individual de progresión del sprint en JIRA. (Fuente: JIRA software)

Inicialmente (antes de febrero 2020), cada equipo de desarrolladores funcionaba por separado, tenían una planificación propia. En realidad, muchas tareas necesitan una coordinación *front-end* y *back-end*, como el tamaño del equipo iba creciendo se decidió cambiar de estructura y agrupar temporalmente *front* y *back-end* para formar equipos en función de la parte de la plataforma en la cual los desarrolladores estaban trabajando.

### La gestión de versiones

En programación en equipo, es muy difícil perderse con las versiones de un código, y dado que pequeñas líneas y modificaciones pueden tener un efecto crítico sobre el resultado de una ejecución es muy importante poder seguir siempre cuales han sido las modificaciones críticas, por quién y cuándo para poder corregirlas. Por eso existe Git, un software de gestión de las versiones muy usado en el mundo del desarrollo por su simplicidad de uso y su comunidad muy activa en GitHub entre otros. Git es un programa gratuito que se puede descargar y ejecutar en la terminal de comandos de su ordenador, pero para un resultado más visual y una gestión más intuitiva se suele usar otro programa con interfaz de usuario. En Datamaran se usa BitBucket, para que sea más cómodo editar una versión de código y compararla con la versión actual (ubicado en la nube y modificada constantemente por los otros miembros del equipo) del código y decidir para cada diferencia cuál versión del código debe de ser la buena. Se basa en el concepto de “ramas”, cuando se quiere trabajar localmente sobre el código, se crea una nueva rama (duplicando el código) para poder editarlo, mientras tanto los otros miembros pueden hacer igual.

Una vez acabado de editar, se sube el código a la nube y se hace una *pull request*, o permisión de juntar el contenido de su rama con la rama principal. En esta etapa, se anuncia a los otros miembros que se pretende juntar la rama para que revisen el código en busca de errores, o problemas antes de validar.



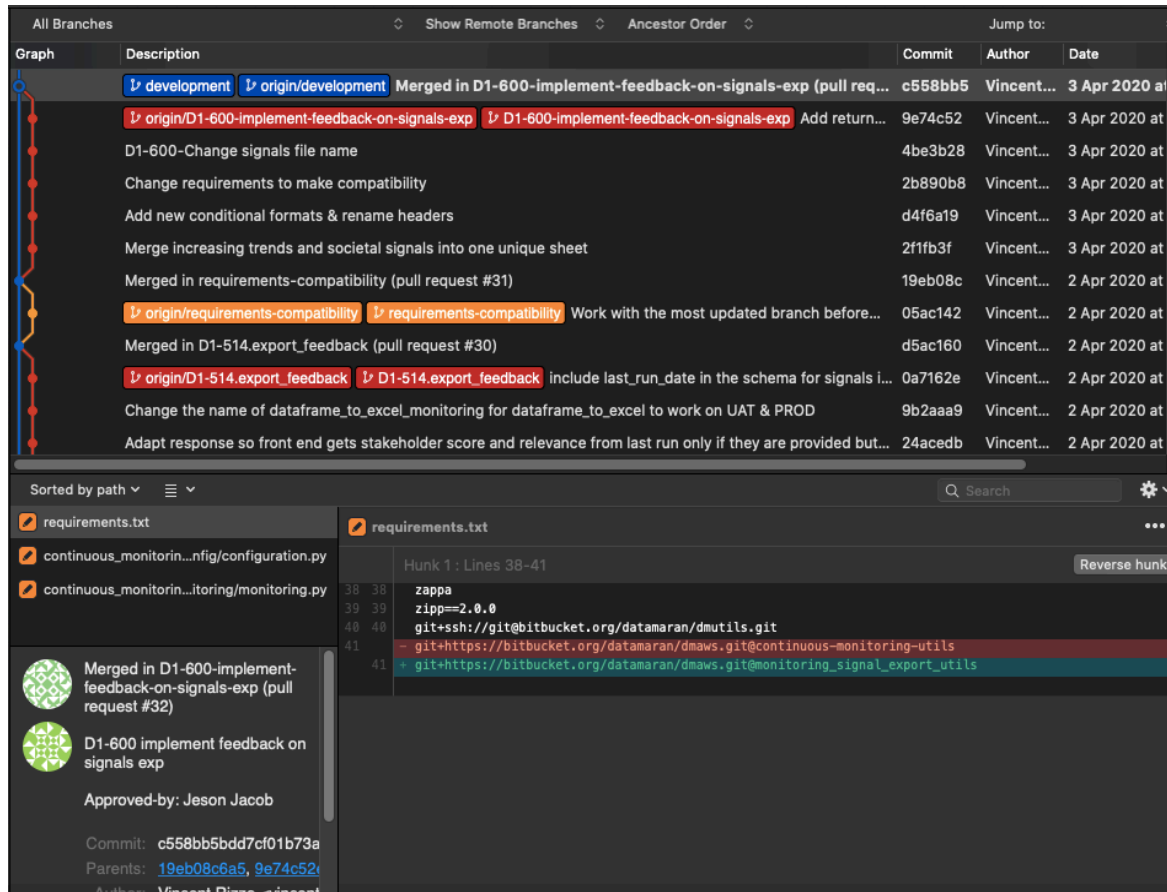


Ilustración 18 - Ejemplo de gestión de versiones en BitBucket gracias al concepto de ramas. (Fuente: BitBucket-SourceTree)

En la Ilustración 18 se observa el funcionamiento de las versiones, cada línea siendo un *commit*, o modificación aportada al código. En cada modificación, abajo se comparan los archivos, subrayando las líneas que han cambiado para facilitar la toma de decisión a la hora de juntar ramas, por ejemplo. Se ve que la rama roja ha ido cambiando en paralelo a la rama azul hasta una *pull request* que ha permitido comparar y juntar las dos ramas.

### Los diferentes entornos<sup>7</sup>

Una vez las tareas realizadas y el código escrito correctamente según el desarrollador quién lo ha escrito, el mismo tiene que pasar a través de un proceso de calidad (*Quality Assurance -QA*) antes de desplegarse para los clientes.

Primero, cuando un desarrollador trabaja, está escribiendo un código sobre su ordenador, si hay que testear el funcionamiento en una página web, se puede usar el mismo ordenador como servidor – pero nadie más tendrá acceso al código. Se dice que se está desarrollando en el entorno local. Una buena práctica de desarrollador es escribir los *unit tests*, se trata de pequeñas funciones que testean el funcionamiento básico del código escrito, si luego se modifica el código, estos *tests* se ejecutarán automáticamente para verificar que la modificación no entra en conflicto con las funcionalidades anteriores.

<sup>7</sup> Entorno en programación define el “escenario” (conjunto de variables y funciones) dónde se está ejecutando un código

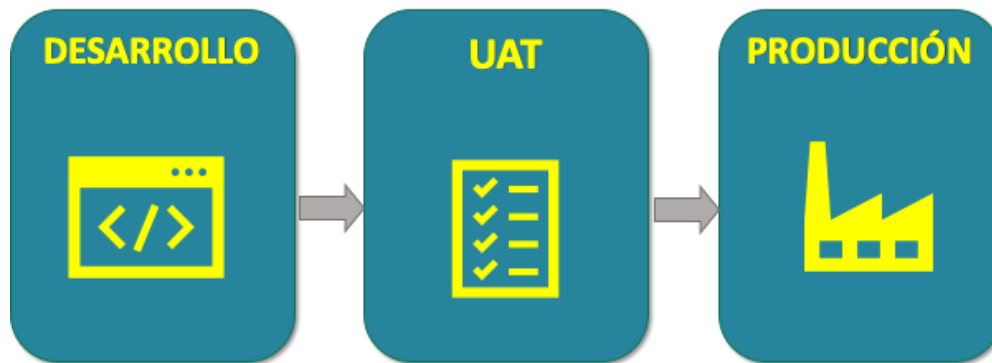


Ilustración 19 - Presentación de la evolución del entorno durante el proceso de calidad para desplegar una aplicación web. (Fuente: elaboración propia).

Cómo se ha detallado en la Ilustración 19, una vez el código ha sido testeado localmente por el desarrollador, se puede desplegar al primer entorno: *development* (o desarrollo), que constituye una primera etapa al despliegue del *software*. El entorno de desarrollo técnicamente está accesible al público que tiene las claves de acceso, pero nunca se facilitan estas claves, se destinan únicamente al equipo de desarrollo para testear el buen funcionamiento del código en el entorno. En desarrollo, el código solo comunica con las librerías<sup>8</sup> que se han definido como necesarias, el resto de la plataforma no existe en este entorno.

Luego se despliega en el entorno *User Acceptance Testing (UAT)*, en este entorno, ya existe una plataforma completa accesible para cualquier usuario del equipo técnico, o de calidad. Este entorno permite testear las funcionalidades en una situación donde se puede usar la plataforma. En UAT, los datos no corresponden a datos reales de usuarios, se hace una diferencia para evitar problemas con la base de datos. Se usa la misma estructura de datos que en la plataforma real, para que todas las cosas queden iguales a parte de los datos.

Finalmente, cuando todo funciona correctamente en UAT, se despliega el código en el entorno de producción, que será el definitivo dónde los clientes tienen acceso y dónde no se pueden aceptar errores, bugs o cosas sin testear.

## 2.6. Conclusión

Datamaran Evalue SLU es una empresa de *Software-as-a-Service* que provee una plataforma llamada Datamaran- Aquella plataforma, destinada a empresas multinacionales de gran tamaño en general, permite realizar un análisis cuantitativo de la posición actual de la política ética de aquella empresa con respecto a la importancia externa atribuida por diferentes *stakeholders* importantes. Entre estos *stakeholders* se toma en cuenta: los competidores de la empresa a través de sus informes corporativos, las leyes y recomendaciones legales, y la opinión pública a través de los artículos de prensa y de Twitter. El algoritmo de Datamaran, basado en el concepto de redes neuronales está entrenado para leer y extraer el sentido profundo de las palabras de manera a ser capaz de ser mejor que un humano a la hora de leer y sintetizar rápidamente una gran cantidad de texto.

El equipo de desarrollo tecnológico de Datamaran se divide entre los *front-end* (se encargan de la experiencia de usuario cuando está navegando en la aplicación) y el equipo *back-end* (se

---

<sup>8</sup> Una librería en programación corresponde a un conjunto de funciones escritas por otros desarrolladores que se pueden importar y usar en otro archivo de código.





encarga de facilitar los datos y su interpretación a la parte *front-end* y los usuarios). Para desarrollar nuevas funcionalidades en la plataforma hay que definir *user stories* en JIRA que luego serán atribuidas a los desarrolladores que irán implementando el código antes de pasar a través de un proceso de calidad para poder ser publicado y utilizable por los usuarios.

Para la mayoría de las comunicaciones internas se usa Slack y Gmail, pero dentro del equipo de desarrollo se planifican las semanas según *sprints* de realización de tareas que se van implementando localmente antes de actualizarse en BitBucket, el programa de gestión de las versiones del código.

## 3. Análisis de la situación actual

### 3.1. Introducción

Datamaran es una empresa que provee una solución informática en línea para gestionar los riesgos acerca de las temáticas sociales y de sostenibilidad para las multinacionales. Su módulo principal de materialidad es un gráfico que ubica las diferentes temáticas dependiendo de su relevancia en el entorno del cliente. Por ello se basa en una serie de medidas cuantitativas de la importancia de las temáticas, usando inteligencia artificial para analizar los textos disponibles en internet.

En este capítulo se pretende explicar en detalles el funcionamiento de la plataforma antes de valorar posibles alternativas de mejoras. Es decir, su estado antes del 22 de abril 2020, día de lanzamiento de la primera fase del proyecto de reforma. Se explicará más específicamente el funcionamiento del módulo de *Continuous monitoring*: el objeto de este trabajo. No obstante, se necesita detallar también el funcionamiento del módulo general *Materiality*, ya que *Continuous monitoring* es parte de ello.

Primero, se detalla el proceso de construcción de un análisis de materialidad, tanto para el cliente como para la infraestructura técnica. Luego se tratará de entender el módulo de Continuous monitoring, objeto de este trabajo. Por último, se realizará un análisis crítico de los diferentes problemas identificados en este módulo.

### 3.2. Desarrollo de un análisis de materialidad

Dentro de las diferentes pestañas accesibles en la plataforma, se encuentra la que se llama *Materiality*. Aquella corresponde a la síntesis y la puesta en perspectiva de la situación de las temáticas en el entorno exterior de la empresa, con la asignación de las prioridades internamente. Dicho de otra manera, se confrontan los recursos que la empresa cliente piensa atribuir a una temática dada con la prioridad que se le debería de asignar dado el contexto actual.

#### 3.2.1. Resultados visibles en *Matrix View*

##### Gráfico de relevancia

El principal objetivo de esta visualización gráfica es entender de manera cualitativa cuáles son las temáticas más relevantes para mi empresa. Por eso se pone en perspectiva para cada temática su importancia interna: se clasifican las temáticas de más importante a menos importante de derecha a izquierda (eje horizontal) y su importancia externa: el software calcula según los parámetros de enfoque, fuentes elegidas el nivel de importancia de la temática para sus competidores, sus clientes y el público en general, de arriba hacia abajo (eje vertical). Las zonas de color en la gráfica permiten identificar muy rápidamente las temáticas más importantes (fondo azul oscuro) de las menos importantes (fondo azul claro).

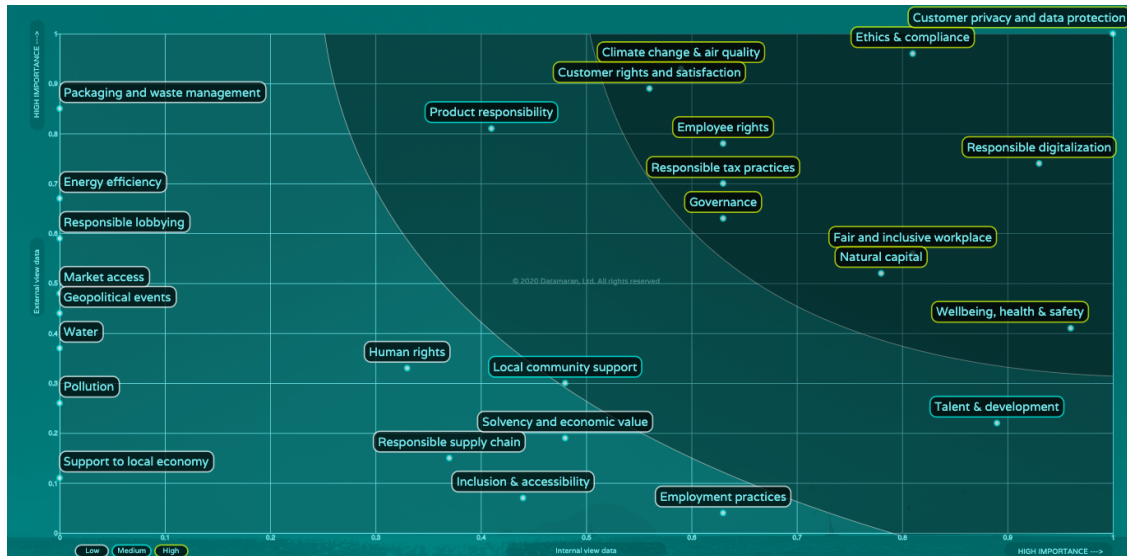


Ilustración 20 - Ejemplo de análisis de materialidad. Verticalmente: importancia externa. Horizontalmente: Importancia interna. (Fuente: Datamaran)

La Ilustración 20 presenta un ejemplo de este análisis, el color de las diferentes zonas (y del contorno del texto definiendo las temáticas) cuantifica la relevancia de la temática para la empresa. Idealmente, se recomienda tener una distribución diagonal: la importancia interna corresponde exactamente a la importancia externa. Existen dos razones principales para que esta alineación no se cumpla:

1. **Inconscientemente**, la empresa ha evaluado diferentemente la importancia externa de una temática (ej. en la Ilustración 20, *Climate change & air Quality* se ubica por encima de la diagonal, demostrando que el cliente subestima su importancia externa). En este caso, la herramienta permite darse cuenta de esta falta de alineación de las medidas tomadas con respecto al riesgo potencial representado por la temática.
2. **Conscientemente**, el cliente estima que una cierta temática es/no es suficientemente relevante para alinear sus medidas con el entorno exterior. Por ejemplo, cómo se ve en la Ilustración 21, *Amazon* da mucha importancia a sus emisiones de dióxido de carbono, el diseño de sus productos y al respeto de los derechos de sus proveedores. Si se comprometen públicamente a respetar estos valores, es normal que, en su matriz de materialidad, valoren internamente más estas temáticas, incluso si en su entorno exterior se ven atribuidas menos importancia.



### Operaciones sostenibles

Adoptamos un método integral y científico de medición y reducción de las emisiones de dióxido de carbono en nuestras operaciones.

[CONTINUAR LEYENDO](#)



### Productos y embalaje

Redesigning packaging to reduce waste and making our private-label products in responsible ways.

[CONTINUAR LEYENDO](#)



### Responsabilidad social

En Amazon estamos absolutamente comprometidos a llevar a cabo nuestras operaciones de manera ética y legal. Ello incluye trabajar con proveedores que respeten los derechos humanos, ofrezcan espacios de trabajo seguros e inclusivos y promuevan un futuro sostenible.

[CONTINUAR LEYENDO](#)

*Ilustración 21 - Algunos valores de Amazon. (Fuente: Amazon Spain)*

Gracias a esta disposición visual, esta parte del software es clave porque contiene la información a más alto nivel, que ayuda visualizar rápida y claramente dónde se ubica una temática y cuál es su nivel actual de riesgo. Es importante decir que Datamaran no provee consejos de ninguna forma a sus clientes, sino más bien que transforma los datos textuales en gran cantidad en un resumen visual muy sencillo tratando de perder la mínima información posible.

### Matriz de materialidad

Para enseñar el gráfico expuesto en el apartado anterior, el software se basa en números concretos proviniendo de la matriz de materialidad. Aquella es una tabla con una línea por cada temática considerada y varias columnas que se van a presentar en este apartado. Primero, la primera columna completamente a la derecha de la tabla corresponde a la relevancia de la temática, es decir en qué zona del gráfico de la Ilustración 20 se encuentra. Esta medida combina la vista interna y externa para entender rápidamente, con una simple escala de 3 niveles cuáles son las temáticas prioritarias en el entorno del cliente. Segundo, la importancia interna (eje horizontal del gráfico Ilustración 20), corresponde al nivel de prioridad de la temática para el cliente, en términos de recursos alojados.

Matrix table ABSOLUTE RANK PERCENTILE RANK UPDATE ANALYSIS

Weight level percentage: 35 30 20 10 5 100 %

Issue Name	Benchmark	M	V	external	external	internal	relevance	
Climate change & air quality	8	7	1	9	2	3	12	High
Customer privacy and data pr...	1	1	3	1	4	1	1	High
Customer rights and satisfacti...	6	3	17	6	10	4	13	High
Employee rights	7	12	10	8	18	7	11	High
Employment practices	25	27	25	22	24	27	11	Low
Energy efficiency	15	14	6	14	11	10	27	Low
Ethics & compliance	2	5	2	7	6	2	6	High
Fair and inclusive workplace	19	16	4	18	9	13	6	High
Geopolitical events	5	26	27	3	1	16	27	Low
Governance	11	18	9	16	16	11	11	High
Human rights	23	17	11	11	7	19	19	Low
Inclusion & accessibility	26	21	24	13	15	26	16	Low
Local community support	14	24	21	12	19	20	15	Medium
Market access	19	8	26	5	5	15	27	Low
Natural capital	23	9	7	20	13	14	7	High
Packaging and waste manage...	12	4	5	17	20	5	27	Low
Pollution	27	6	19	27	14	21	27	Low
Product responsibility	11	2	14	10	23	6	17	Medium
Responsible digitalization	3	19	15	2	3	8	3	High
Responsible lobbying	21	10	13	4	8	12	27	Low
Responsible supply chain	19	23	20	25	21	24	18	Low
Responsible tax practices	5	13	18	15	12	9	11	High
Solvency and economic value	9	25	23	24	27	23	15	Low
Support to local economy	24	20	17	21	25	25	27	Low
Talent & development	13	22	23	19	26	22	4	Medium
Water	19	15	8	26	17	18	27	Low
Wellbeing, health & safety	20	11	13	23	22	17	2	High

Ilustración 22 - Ejemplo de matriz de rangos materialidad. Columnas de izquierda a derecha: nombre de la temática, Benchmark, Regulaciones obligatorias, Recomendaciones legales, Noticias, Redes sociales (Twitter), Score externo, Score externo y Relevancia. (Fuente: Datamaran)

En la Ilustración 22 la columna con fondo azul diferente de las demás representa la importancia interna, los números son rangos de 1 a 27 (porque este análisis trata 27 temáticas, pero depende de la configuración de usuario).

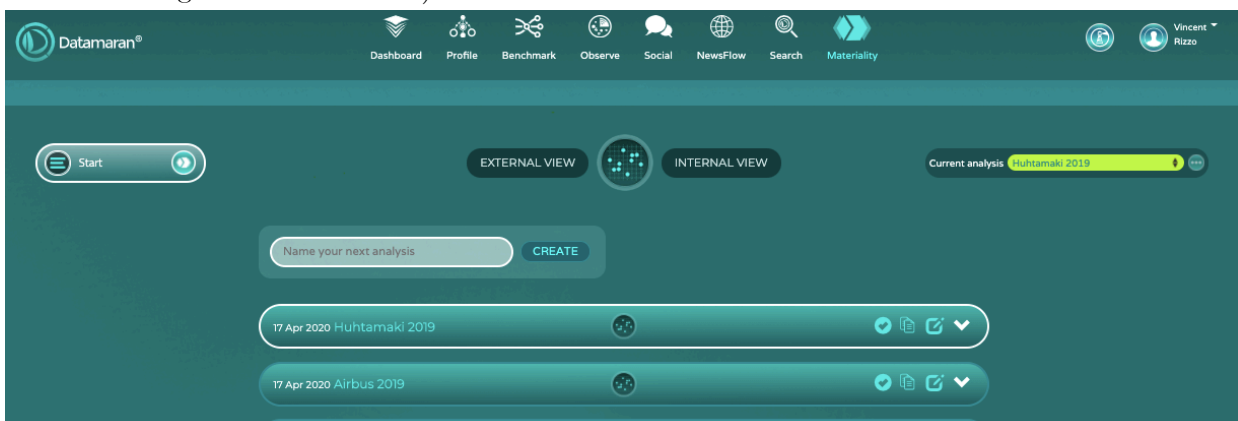


Ilustración 23 - Página de selección de un análisis por el usuario en el módulo de materialidad. (Fuente: Datamaran)

Esta cuantificación se hace en la parte *Internal View* accesible desde la página base de materialidad (Ilustración 23), esta página se ve en la Ilustración 24 . En esta página se puede usar la interfaz para atribuir una importancia relativa a cada temática. La primera columna a la derecha (*Internal View score*) muestra lo que se llama el *percentile rank*<sup>9</sup> interno de la temática.

<sup>9</sup> El percentile rank de un valor dentro de una serie corresponde a un número real entre 0 y 1 representando el porcentaje de valores de la serie que se encuentra por debajo de dicho valor.



Ilustración 24 - Vista interna de parametrización de la matriz. (Fuente: Datamaran)

Se emplea el *percentile rank* para los cálculos, y el ranking bruto para enseñar los datos al usuario. Se usa el rango cuando se comunica información bruta al usuario, porque el número es más intuitivo de entender (ej. La temática **Cambio climático** está en 4º puesto), pero para el gráfico, o para los cálculos es necesario tener una medida que tiene ciertas propiedades importantes:

1. **Normalizado:** el *percentile rank* siempre está entre 0 y 1, independientemente del número total de temáticas considerado.
2. **Ascendiendo:** 0.9 es mejor que 0.5 en *percentile rank*, pero 2 es mejor que 5 en ranking absoluto. Permite que en el gráfico lo más relevante esté arriba a la derecha.

En muchas ocasiones el *percentile rank* se calcula:

$$Pc(T_i) = 1 + \frac{r_i - 1}{N} \quad (2)$$

$T_i$  la temática número  $i$  ( $N$  siendo el total de temáticas consideradas).

$r_i$  el puesto de  $T_i$

$Pc(T_i)$  el *percentile rank* de  $T_i$

Cómo se puede ver en la Ilustración 22, la matriz de materialidad se compone de una columna de **Score externo** que, a través del *percentile rank* corresponde a los datos del eje vertical en el gráfico de la Ilustración 20. Este puesto se calcula a base de un medible, llamado el *stakeholder score*, que es una media ponderada de los *percentile rank* de cada fuente<sup>10</sup>. Cómo se puede ver en la Ilustración 22, cada una de las fuentes tiene un peso (porcentaje encima de la columna), atribuido por el usuario. Datamaran afecta por defecto un peso recomendado a cada una de las fuentes:

1. Benchmark: 35% en total. Es posible definir hasta 5 *benchmarks* diferentes (cambiando las empresas observadas por el mismo). Cada uno llevará por defecto una parte igual

<sup>10</sup> Fuente se refiere a una de las áreas de análisis entre *benchmark*, regulaciones obligatorias, recomendaciones legales, noticias de prensa y redes sociales.

- del total de 35% (ej. Si 2 *benchmarks* se definen, llevarán 17,5% en peso cada uno).
2. Regulaciones obligatorias (o leyes): 30%
  3. Recomendaciones legales (o informes escritos por entidades sin poder legislativo real, como ONG, por ejemplo, y constituyen una recomendación): 20%
  4. Noticias (prensa): 10%
  5. Redes sociales (Twitter): 5%

El usuario tiene también la posibilidad de aportar sus propias fuentes de datos, a condición de tener en una hoja Excel una clasificación de las temáticas definidas, se puede añadir hasta 10 fuentes personalizadas de datos. También se deja libre la definición del peso aplicado a la fuente, con lo cual puede pasar, en casos extremos, que las fuentes de Datamaran definidas anteriormente no tengan peso comparable con los datos personalizados.

Entonces el *stakeholder score* se calcula gracias a la fórmula siguiente:

$$Stakeholder\ score(T_i) = \sum_{f \in F} p_f \cdot Pc(T_i, f) = \sum_{f=1}^{N_b + N_p + 4} p_f \cdot Pc(T_i, f) \quad (3)$$

$T_i$  la temática número  $i$  ( $N$  siendo el total de temáticas consideradas).

$f$  es el número de la fuente considerada, que pertenece al conjunto  $F$  de índices de las fuentes.

$N_b$  es el número de *benchmarks* declarados.

$N_p$  es el número de fuentes personalizadas aportadas por el usuario.

$p_f$  el peso normalizado de la fuente de índice  $f$  (ej. si se atribuye 5% a Twitter,  $p_{Twitter}$  será 0,05).

$Pc(T_i, f)$  es el *percentile rank* de  $T_i$  para la fuente  $f$ , asociado al puesto de esta temática dentro la fuente  $f$ .

En resumen, en cada una fuente, cada temática tiene un puesto atribuido que refleje su importancia, este puesto se convierte en un número entre 0 y 1 como en la ecuación (2) (el *percentile rank*). Luego, con los pesos atribuidos por el usuario a cada fuente, y los respectivos *percentile rank* se calcula el *stakeholder score* según la ecuación (3), una vez clasificados todos los *stakeholder scores* para cada temática, se obtiene el puesto general de la temática, lo que, una vez convertido en *percentile rank* permitirá ubicar la temática en el gráfico de relevancia de la Ilustración 20.



### 3.2.2. Parametrización y cálculo de las fuentes

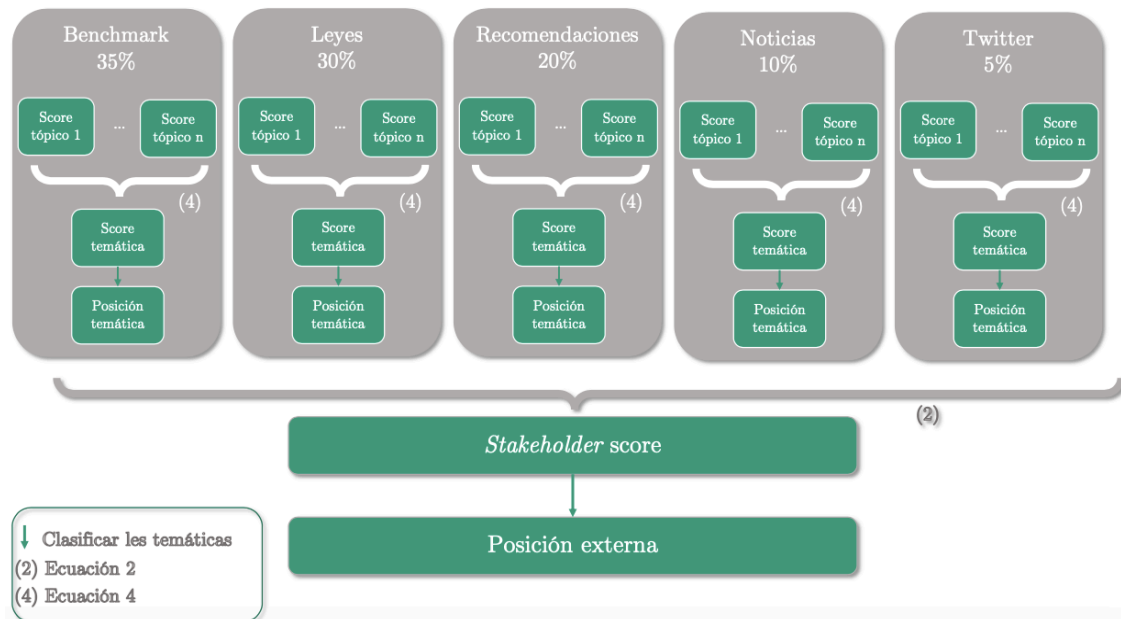


Ilustración 25 - Esquema resumen de los cálculos realizados para evaluar la importancia externa de una temática. (Fuente: elaboración propia)

En la Ilustración 25 se sintetiza el proceso general de cálculo del ranking “externo” de una temática. Se ve cómo se calcula primero un ranking en cada una de las fuentes, antes de pasar a calcular el *stakeholder score* (una media ponderada de los rankings por fuente) y clasificar las temáticas usando esta medida. A continuación, se detalla más precisamente las diferentes etapas de este proceso.

Antes de poder obtener el gráfico de resultados como en la Ilustración 20, es obligatorio ajustar los diferentes parámetros del análisis. En primer lugar, es importante decir que las temáticas no son las mismas para todos los usuarios: los usuarios pueden cambiar los nombres de aquellas y definir más precisamente lo que contienen.

Datamaran define por defecto una ontología, es decir la definición de cada tópico, que conceptos contiene exactamente. El Anejo D muestra cómo cada temática se divide en tópicos específicos, aquellos tópicos contienen varias palabras claves que los expertos lingüísticos de Datamaran pueden, si procede, cambiar a medida que evolucionan las necesidades actuales. Lo que se deja personalizable para el usuario es:

1. Definir los tópicos que encajan en una temática dada y los pesos relativos correspondiente gracias al mapa de tópicos.



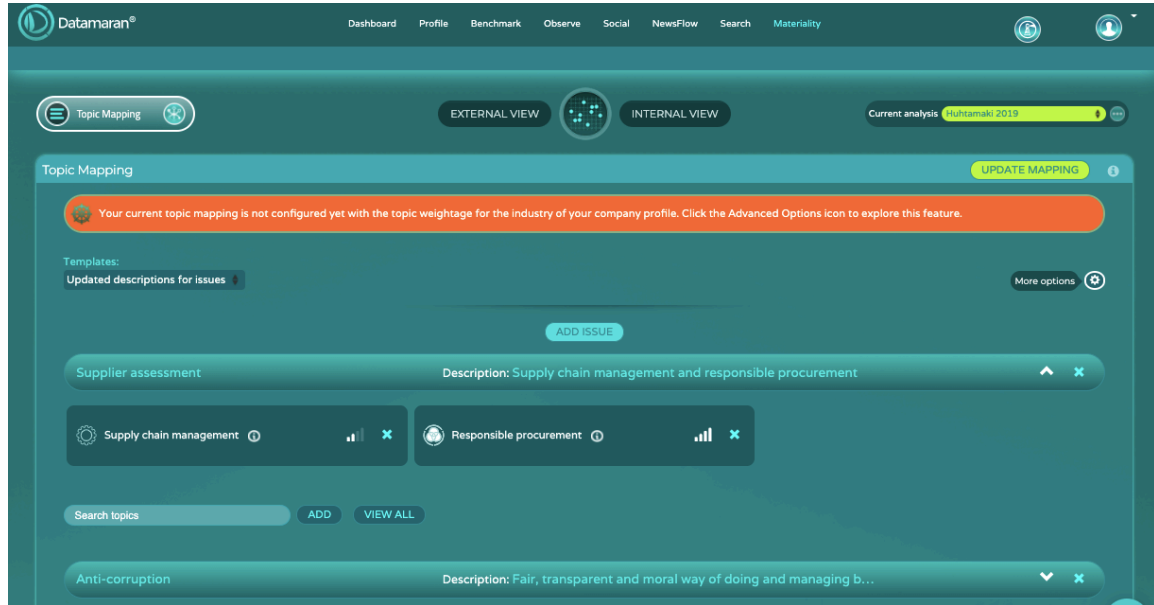


Ilustración 26 - Ejemplo de mapa de tópicos para la temática Supplier Management. (Fuente: Datamaran)

La Ilustración 26 muestra la página de mapa de tópicos, como un usuario puede seleccionar qué tópicos entran en una temática, y que peso relativo tienen aquellos.

$$p(t_j) = \frac{w(t_j)}{\sum_{k \in T} w(t_k)} \quad (4)$$

$t_j$  es el tópico número  $j$  dentro del conjunto  $T$  incluyendo todos los tópicos definidos.

$p(t_j)$  es el peso global de  $t_j$

$w(t_j)$  es el peso local de  $t_j$ , se define gracias al símbolo parecido a barras de señal Wi-Fi que puede variar de 0,25 a 1.

La ecuación (4) muestra cómo se calculan los pesos globales de los tópicos a nivel de una temática, a partir de los pesos locales atribuidos por el usuario.

2. Agregar y definir nuevos tópicos desde 0 gracias a la pestaña de *Community topics*. Estos tópicos personalizados exponen el análisis al riesgo de no tener datos fiables en ciertas temáticas, ya que no se puede controlar perfectamente la validez de los tópicos definidos por el usuario (los tópicos de la ontología de Datamaran se validan por expertos lingüísticos y a través varios análisis de cobertura).

Personalizar libremente el nombre de una temática. Lo que implica que los nombres de temáticas serán, a priori, diferentes para cada análisis.

Por la división de las temáticas en tópicos, para cada fuente en la cual se calcula un dado *score* del tópico, se calculará la media ponderada de los *scores* de los tópicos pertenecientes a la temática con sus respectivos pesos:

$$S(T_i) = \sum_{j \in T_i} p(t_j) \cdot S(t_j) \quad (5)$$

$T_i$  es la temática número  $i$  dentro del conjunto  $I$  incluyendo todas las temáticas definidas. Corresponde al conjunto de índices de los tópicos incluidos en la definición de la temática.

$t_j$  es el tópico número  $j$  dentro del conjunto  $T$  incluyendo todos los tópicos definidos.

$p(t_j)$  es el peso de  $t_j$  dentro de  $T_i$ , de acuerdo con la ecuación (4).

$S(t_j)$  es el *score* del tópico  $t_j$  para una fuente dada.

$S(T_i)$  es el *score* de la temática  $T_i$  para una fuente dada.

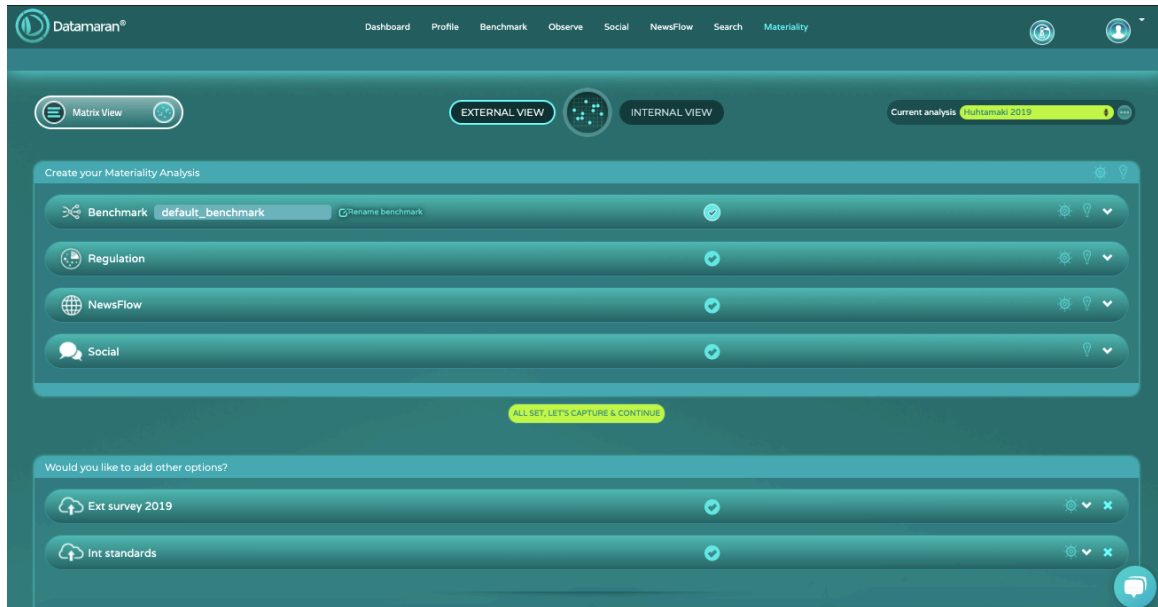


Ilustración 27 - Pestaña Materiality/External view, dónde se entran los parámetros del enfoque para cada fuente. (Fuente: Datamaran)

Para poder definir el enfoque de cada fuente, es necesario ir a la página *External view*, que se detalla en la Ilustración 27. Para cada fuente se detallan los diferentes parámetros modificables por el usuario. Cada una de estas fuentes tiene una pestaña dedicada a nivel general, donde se puede experimentar para ver en concreto cuáles son los resultados que se obtendrían aplicando una cierta configuración.

Antes que nada, el cliente tiene que definir su perfil, es decir a qué empresa pertenece, introduciendo parámetros básicos como su tamaño, países de actividad, de proveedores entre otros.

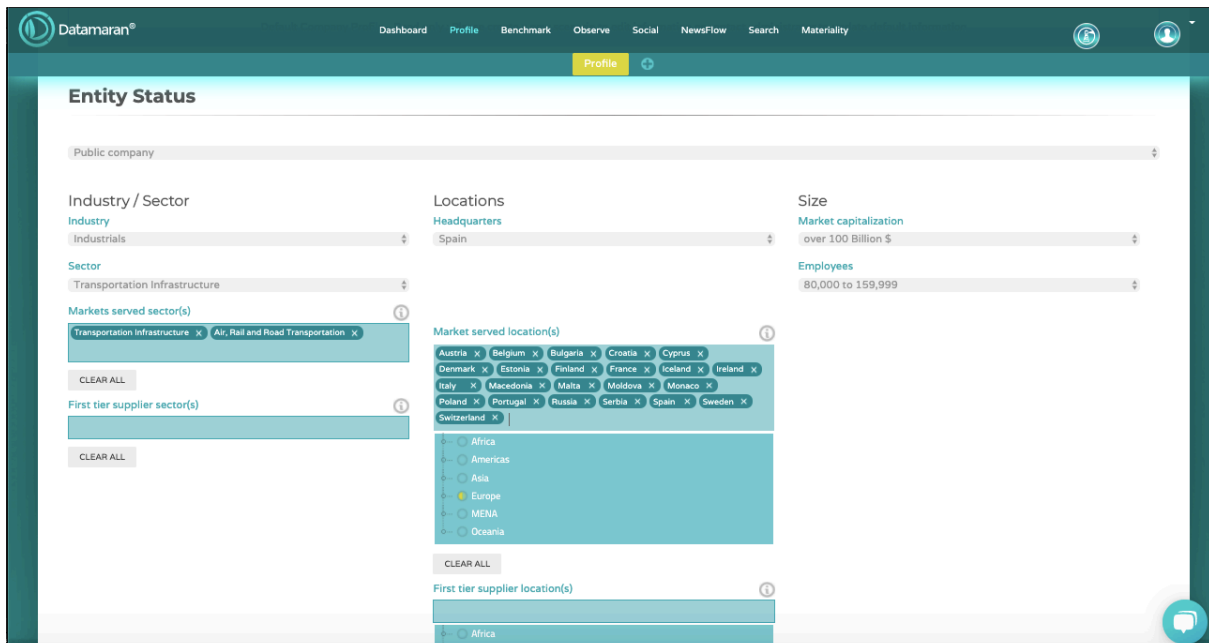


Ilustración 28 - Pestaña de definición del perfil de usuario. (Fuente: Datamaran)

Los datos introducidos en los campos de la pestaña perfil (Ilustración 28) permiten sobre todo definir los *peers* de la empresa cliente.

Una vez correctamente reseñado el perfil se puede empezar examinar cuáles son los datos obtenidos en la base de datos de Datamaran cuando se aplica cierta configuración, y eso en cada

pestaña correspondiente a cada tipo de fuente.

### 3.2.3. Parámetros del Benchmark

Esta fuente se encarga de cuantificar en los informes oficiales (públicos) de las empresas teniendo una actividad comparable a la del cliente cuáles son las temáticas más comentadas. Para ello se pueden usar los informes financieros (10-K), los informes de sostenibilidad o los informes de la SEC<sup>11</sup> que tratan de las empresas determinadas previamente.

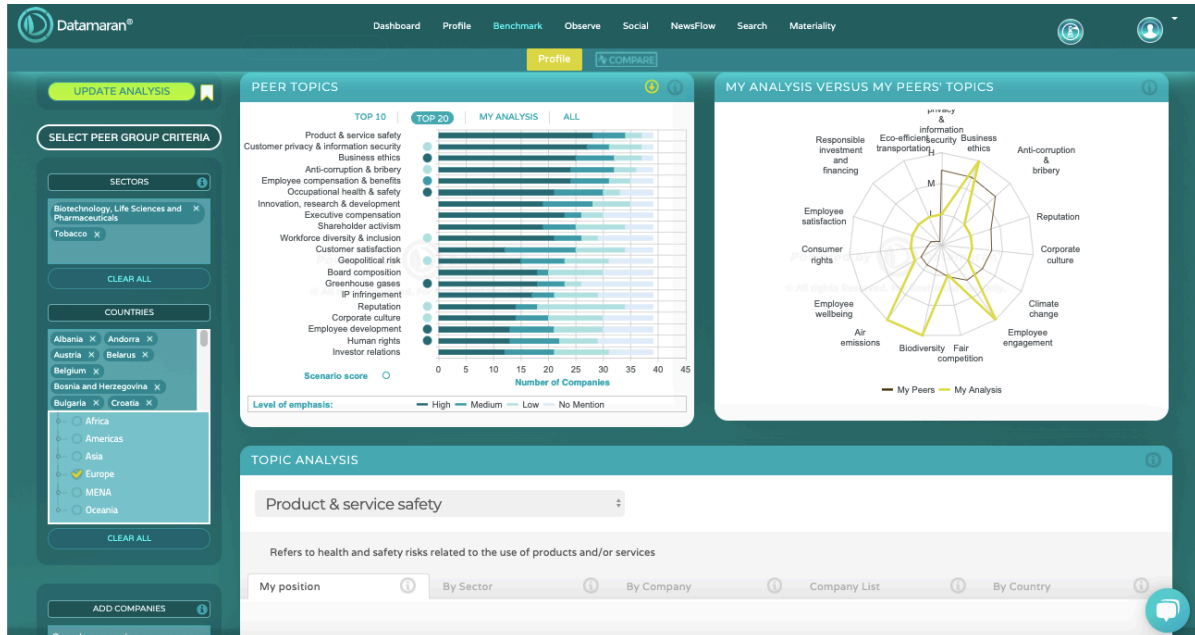


Ilustración 29 - Pestaña Benchmark libre para la exploración de los informes de los competidores. (Fuente: Datamaran)

Cuando se accede a la pestaña benchmark, se puede ver una página como en la Ilustración 29. A la izquierda se dispone de varios campos para introducir las informaciones necesarias a la definición del enfoque del análisis. La Ilustración 30 muestra todos estos campos:

<sup>11</sup> La Securities and Exchange Commission en los Estados Unidos es la entidad no gubernamental que se encarga de verificar el cumplimiento de las leyes de los mercados financieros.



Ilustración 30 - Parámetros modificables por el usuario en la pestaña de benchmark. (Fuente: Datamaran)

1. Sectores: aquí se introducen los sectores de actividad económica de la empresa, se considerarán automáticamente todas las empresas teniendo actividad en estos sectores. En el ejemplo de la Ilustración 30 se seleccionarán todas las empresas que declaran una actividad en el negocio de venta comida y bebidas o de todos productos de consumo.
2. Países: en este campo se restringe los países de ubicación de la sede principal de las empresas seleccionadas por la restricción de sectores. En el ejemplo de la Ilustración 30, una vez se hayan seleccionado todas las empresas pertenecientes a uno de los dos sectores, se filtró esta lista guardando únicamente las empresas que tienen su sede principal en un país europeo.
3. Compañías: se puede entrar el nombre de una o varias compañías en particular (existe un sistema de sugestión automática que propone nombres solo con entrar las primeras letras).
4. Listas de competidores: una vez se han registrado varias compañías en el campo anterior, se puede guardar la lista como lista de compañías similares para facilitar su agregación posterior.
5. Fuentes de datos: filtra por tipos de informes que se deben analizar (financieros, de sostenibilidad o de la SEC).
6. Año: año de publicación de los informes considerados, por defecto será el año anterior.

Los campos compañías y lista de competidores no son restrictivos, es decir que se tomarán en cuenta todas las empresas que están al menos en una de las listas, sin obligatoriamente estar en todas.

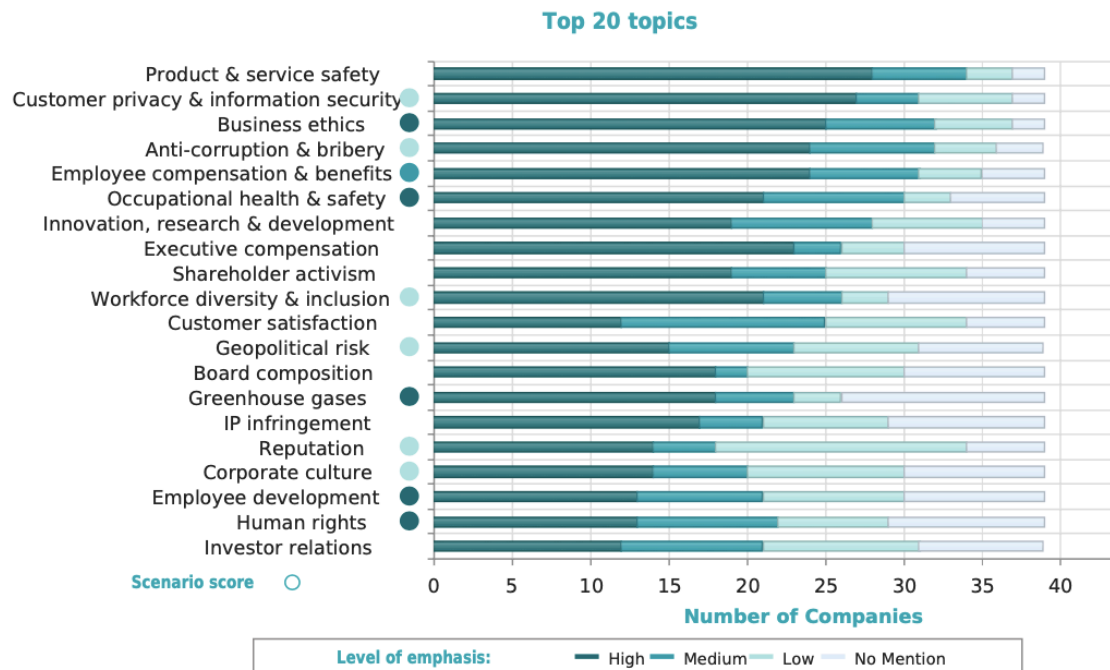


Ilustración 31 - gráfico de nivel de mención por tópico en los informes anuales de las empresas de tamaño grande del sector agroalimentario en Europa. (Fuente: Datamaran)

Una vez aplicados todos estos filtros, se puede observar los resultados como en la Ilustración 31 y la Ilustración 32. La Ilustración 31 muestra el nivel de mención de los 20 tópicos más representativos de los informes seleccionados. La clasificación se efectúa a partir del porcentaje de mención de las palabras del tópico en el texto, o porcentaje de *hit*, calculado según un método confidencial. A la izquierda del gráfico en frente de cada uno de los 20 tópicos se encuentra un círculo de color representando la posición de la empresa cliente en base a sus informes.

De cara a poder clasificar los tópicos, Datamaran define una medida llamada *emphasis* que condensa la información del gráfico de barras de la Ilustración 31:

$$e(t_j) = \frac{3 \cdot H(t_j) + 2 \cdot M(t_j) + L(t_j)}{H(t_j) + M(t_j) + L(t_j) + NM(t_j)} = \frac{3 \cdot H(t_j) + 2 \cdot M(t_j) + L(t_j)}{N_{\text{compañías}}} \quad (6)$$

$t_j$  es el tópico número  $j$  dentro del conjunto  $T$  incluyendo todos los tópicos definidos.

$e(t_j)$  es la *emphasis* de  $t_j$ .

$H(t_j)$  es el número de informes que mencionan **mucho** el tópico  $t_j$ : clasificación *High*.

$M(t_j)$  es el número de informes que mencionan **bastante** el tópico  $t_j$ : clasificación *Medium*.

$L(t_j)$  es el número de informes que mencionan **poco** el tópico  $t_j$ : clasificación *Low*.

$NM(t_j)$  es el número de informes que **no mencionan** el tópico  $t_j$ : clasificación *No mention*.

$N_{\text{compañías}}$  corresponde al número total de compañías estudiadas.

Para cada temática (conteniendo varios tópicos con diferentes pesos), se calculará la *emphasis* de aquella cogiendo la media ponderada de las *emphasis* de los tópicos que componen la temática, de acuerdo con la ecuación (5). Será esta *emphasis* a nivel de la temática que permitirá clasificarlas y obtener el percentile rank de aquellas para finalmente agregarlo a su *score* externo.

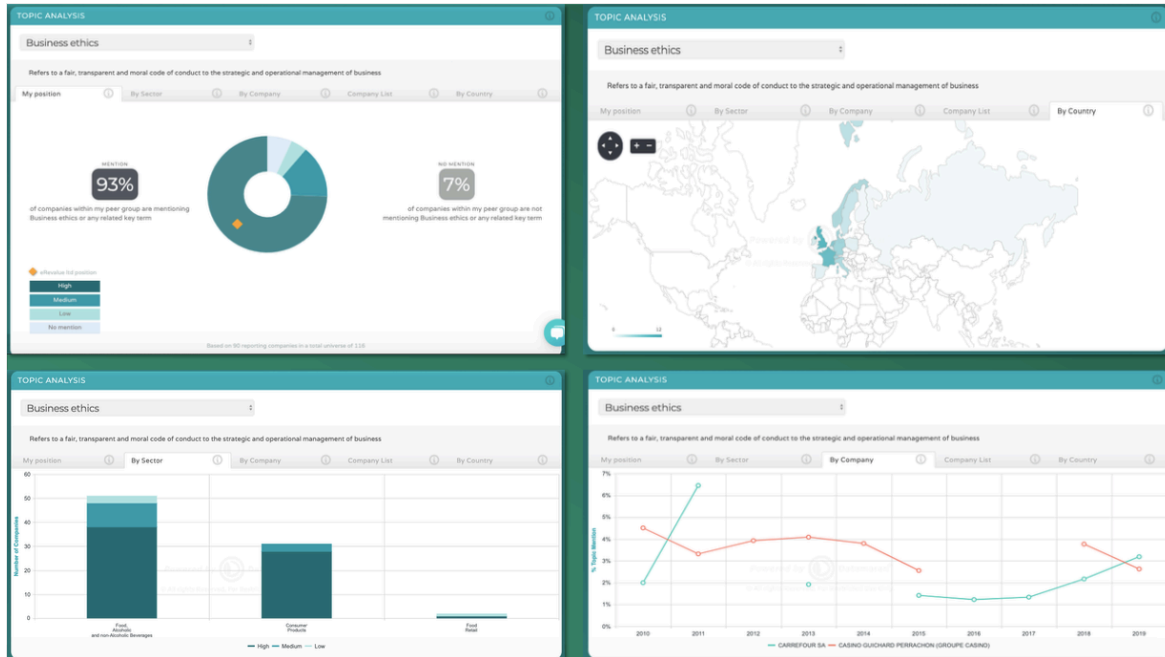


Ilustración 32 - Diferentes gráficos por tópico de la pestaña Benchmark. De arriba hacia abajo y de izquierda a derecha: Posición del usuario, desglose geográfico, desglose por sector de actividad económica, desglose por compañía. (Fuente: Datamaran)

Benchmark no solo sirve para calcular el emphasis, sino que también dispone de varias herramientas gráficas que permiten mejor visualizar los datos que se están añadiendo al análisis. La Ilustración 32 muestra 4 vistas diferentes de los datos de *benchmark*, tomando como ejemplo el tópico *Business ethics*:

1. Posición del usuario: este gráfico permite ver la repartición en porcentaje de los diferentes niveles de mención de un tópico seleccionado, y en qué parte se ubica el usuario para poder observar en detalle cuáles son los tópicos en los cuales más diferencia tiene con sus competidores.
2. Geográfico: dentro de los países que se han seleccionado (aquí Europa solamente), se puede observar que países tienen las compañías que más mencionan el tópico estudiado.
3. Por sector: se desglosan los datos por sector de actividad de las empresas que lo mencionan, para obtener una visión más clara de si el tópico es más importante para un sector que para otro.
4. Por compañía: se muestra para cada compañía introducida en el campo *Companies* de parametrización (véase Ilustración 30) el nivel de *emphasis* de aquella a lo largo del tiempo. Para ciertos años Datamaran no tiene datos porque los informes no se han encontrado todavía.
5. Perfil del usuario: el gráfico de la Ilustración 33 muestra el perfil de la compañía del usuario comparado con el perfil medio de las empresas seleccionadas en el enfoque

### My analysis versus my peers' topics

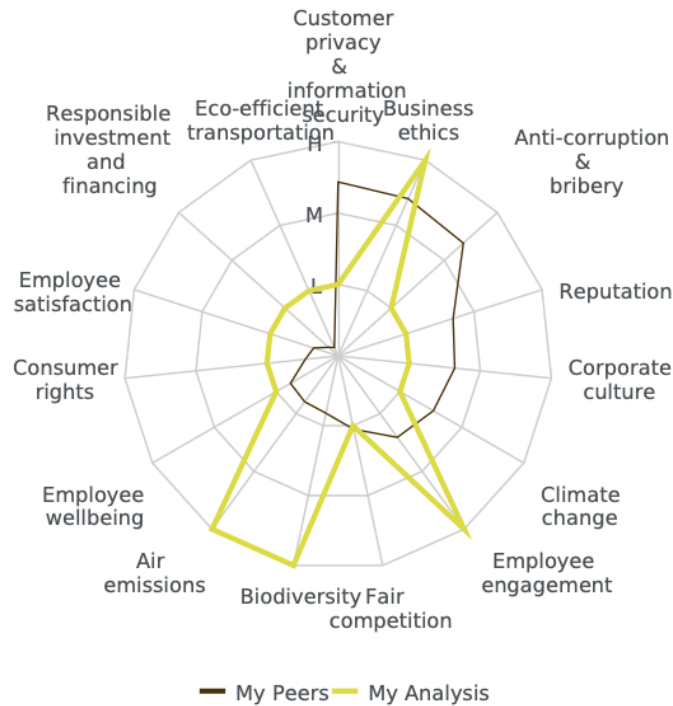


Ilustración 33 - gráfico perfil del usuario en la pestaña benchmark. (Fuente: Datamaran)

Una vez el usuario haya podido testear y observar los resultados que producirían seleccionar un cierto enfoque, ya lo puede entrar definitivamente en la pestaña de materialidad para que se tome en cuenta en su análisis de materialidad.

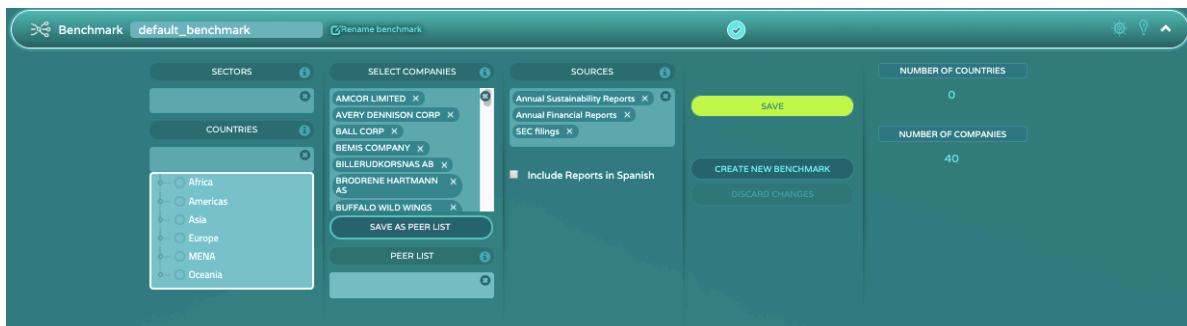


Ilustración 34 - Panel de parametrización de Benchmark en la pestaña de materialidad. (Fuente: Datamaran)

Cómo se ve en la Ilustración 34, el panel de Benchmark, que se obtiene expandiendo la vista de la Ilustración 27 permite introducir los datos finales que se han seleccionado para efectuar el análisis de materialidad por el usuario.

### 3.2.4. Leyes y recomendaciones legales

La pestaña *Observe* de la Ilustración 11 contiene la parametrización para dos fuentes en realidad: regulaciones obligatorias (leyes) y recomendaciones legales. Este módulo busca textos definidos como *regulations* en inglés en la base de datos de Datamaran. Algunos son textos legales emitidos por entidades teniendo poder legislativo, otros simplemente son recomendaciones redactadas por ONGs, otras entidades pero que no tienen valor de ley. Se consideran de manera separada estos dos aspectos de los textos porque no se le puede pretender aplicar el mismo peso a la hora de calcular el *score* final. No obstante, la existencia de una recomendación legal puede ser



una señal de la existencia futura de una ley o normativa que trata de incluir las recomendaciones del primer texto.

Ilustración 35 - Pestaña Observe: formulario de selección del enfoque de las regulaciones. (Fuente: Datamaran)

La vista de la Ilustración 35 muestra los diferentes campos de parametrización de las regulaciones que el usuario introducirá. De todos estos campos podemos destacar:

1. *Enforcement level*: será *Mandatory* o *Voluntary* para diferenciar entre leyes y recomendaciones legales.
2. *Topics of interest*: se puede introducir algunos tópicos porque reduzcan el enfoque de los textos que entrarán en el recuento de regulaciones
3. *Year range*: filtra sobre el año de introducción (cuando haya sido escrita y votada) o el año de entrada en vigor. Por defecto, se coge a partir de 2000, entonces cualquier texto escrito o entrado en vigor durante este periodo se tomará en cuenta. *Updated after* corresponde a todos los cambios en el texto que se pueden hacer posteriormente, ej. si una ley cambia de sector de restricción para aplicarse a toda la industria del transporte en vez de solamente aplicarse a los transportes por ferrocarril.
4. *Geographical scope*: se entra una lista de países para filtrar las regulaciones

La segunda pestaña – *My regulatory radar* – se basa en los datos del perfil de usuario reseñados para crear un *dashboard* con toda la información más relevante actualmente para la compañía del usuario.



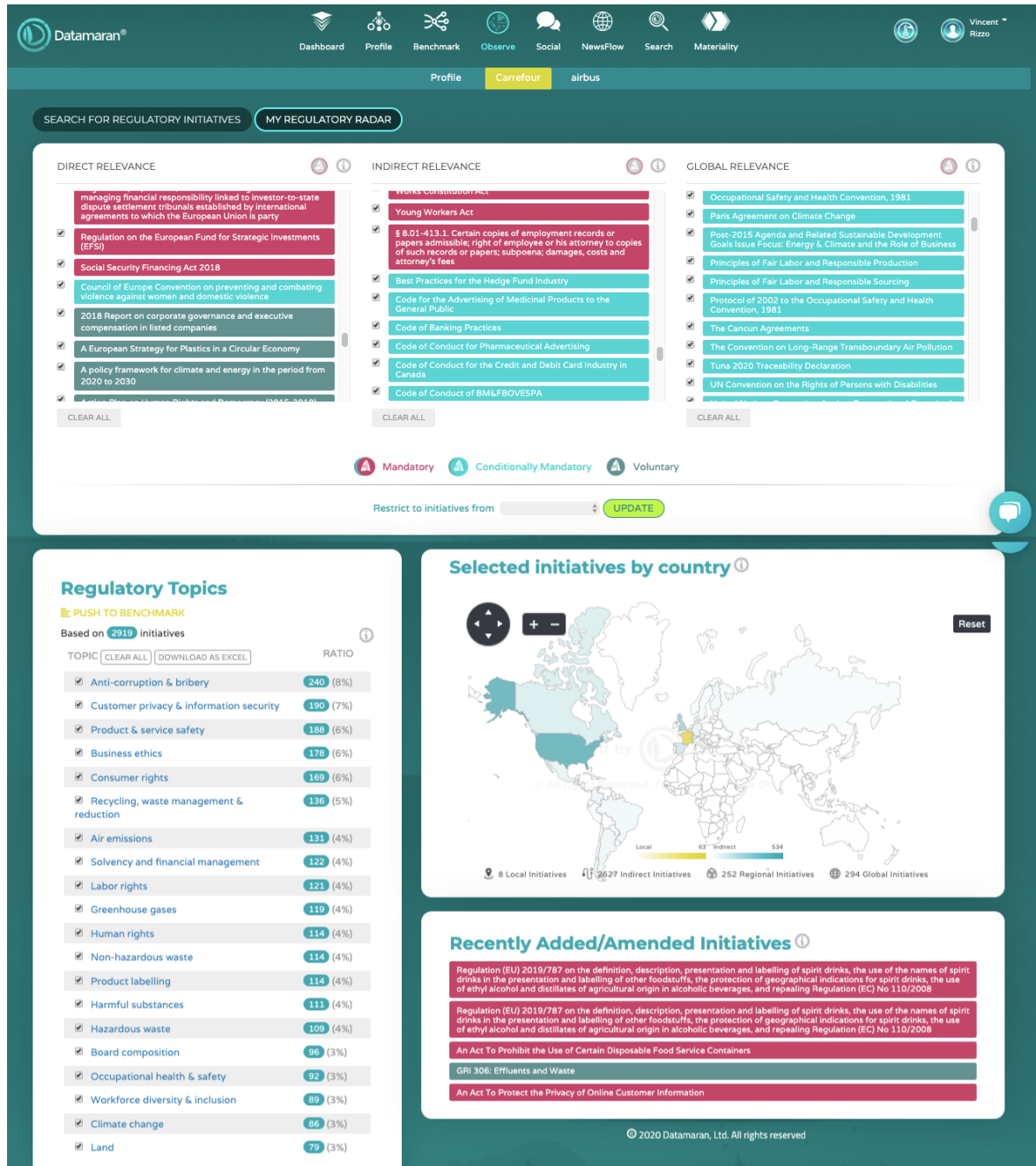


Ilustración 36 - Pestaña My Regulatory Radar. Gráficos de arriba hacia abajo, de izquierda a derecha: lista de regulaciones por tipo de relevancia, tópicos con más regulaciones, regulaciones por país, lista de textos más recientes. (Fuente: Datamaran)

Como lo muestra la Ilustración 36, la pestaña de regulaciones facilita varias herramientas para sintetizar los diferentes textos que interesan el perfil del usuario. Entre otros se clasifican los textos por relevancia:

1. Relevancia directa: el texto hace referencia directa al sector de actividad económica del usuario, y las restricciones se aplican específicamente al sector y el tipo de empresa correspondiente.
2. Relevancia indirecta: el texto viene de un país diferente o puede afectar los proveedores del usuario
3. Relevancia global: se trata de medidas a nivel más alto que el único sector de actividad, como una ley general europea sobre el tratamiento de residuos.

Se propone también una visión rápida de los tópicos más mencionados en regulaciones y con qué porcentaje. En la parte baja y a la derecha se observa un mapa de proveniencia de las regulaciones de la lista, por tipo de relevancia. La lista de los últimos textos publicados muestra las regulaciones relevantes entrada en vigor el mismo mes, dando un clic a alguna de ellas abre una ventana pop up con todos los detalles de la regulación cómo se ve en la Ilustración 38.

Una vez se hayan seleccionado correctamente los filtros para regulaciones, se puede reseñar estos datos en la parte de materialidad para fijar el enfoque del análisis.

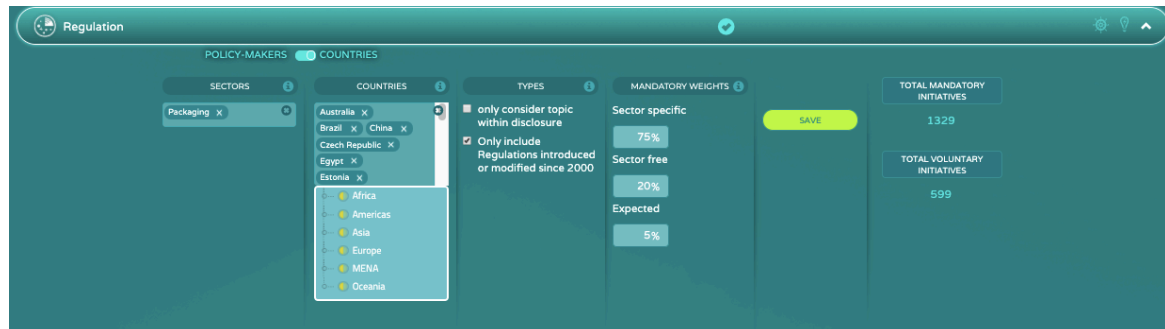


Ilustración 37 - Panel de parámetros para las regulaciones (obligatorias y recomendadas) en la pestaña *Materiality/External view*. (Fuente: Datamaran)

En la Ilustración 37 se muestran los parámetros que se pueden reseñar para las regulaciones, a parte de los ya explicados en la pestaña *Observe* están las opciones:

1. Considerar sólo los tópicos en el alcance<sup>12</sup> de la normativa
2. Incluir solo las regulaciones después de 2000: por defecto se considera que no interesan las normativas que no estén al menos actualizadas después del año 2000. Es importante decir que el periodo de estudio no es de tamaño fijo por esta característica, cada año se toman en cuenta más regulaciones y entonces el volumen total tiende a aumentar.
3. Pesos de las regulaciones obligatorias: los textos considerados obligatorios se dividen en 3 categorías:
  - a. *Sector specific*, los que se dirigen específicamente a ciertos sectores de actividad económica, y dentro de estos sectores se encuentra el sector del usuario (recomendado: 75%).
  - b. *Sector free*: son las regulaciones globales, a nivel nacional o supranacional sin dirigirse a cualquier sector en particular (recomendado: 20%).
  - c. *Expected*: en las dos categorías anteriores solo se cuentan los textos redactados, entrados en vigor o actualizados entre 2000 y el año corriente. Las regulaciones previstas para el futuro, que se plantean para una fecha posterior al día (recomendado: 5%).

<sup>12</sup> Aquí alcance se refiere a los tópicos que son el objeto del texto de la normativa, en oposición a los tópicos que solo se mencionan si ser el objeto de cualquier restricción especial en el texto.

## GRI 306: Effluents and Waste

(3240) [Click to follow](#)

Voluntary DISCLOSURE REQUIREMENTS: Yes

### DESCRIPTION

This Standard sets out reporting requirements on the topic of effluents and waste. It is meant to be used together with the GRI 101: foundation, GRI 103: Management Approach and GRI Standards Glossary documents. The document includes Requirements, recommendations and guidance; any organisation must only comply with the requirements in order to claim that its report has been prepared in accordance with the GRI standards. This Standard encourages reporting on water discharge by quality and destination, waste by type and disposal method, as well as any significant spills that have occurred, any transport of hazardous waste and any impact on water bodies by water discharges and runoff.

### LINKS:

[Link to download the official text in English](#) , [Official text in English](#)

### Related Initiatives

Amended by [GRI 303: Water and Effluents 2018](#)  
Relates to [GRI 101: Foundation](#)

### TOPICS REFERENCED

Within Disclosure Requirements: Spills Water pollution Non-hazardous waste

Outside of Disclosure Requirements: Hazardous waste Biodiversity

### SDGs

SDG2: ZERO HUNGER SDG3: GOOD HEALTH AND WELL-BEING SDG6: CLEAN WATER AND SANITATION SDG14: LIFE BELOW WATER  
SDG15: LIFE ON LAND

### LIFE CYCLE

Introduction: Oct, 19, 2018  
Entry into Force: Jul, 1, 2018  
Last update: Jan, 1, 2021  
Evolution of the initiative: In 2016, the GRI G4 Guidelines transitioned into the GRI Sustainability Reporting Standards (GRI Standards). A public comment period for the draft ended on the 17th of July 2016. On the 5th of October 2016, it was announced by the GRI that the Standards would be released on the 19th of October 2016. This Standard is effective for reports published from the 1st of July 2018 onwards, although earlier adoption is possible and encouraged by the GRI.

### Policy-maker

Name: [Global Reporting Initiative \(GRI\)](#)  
Type: Non-profit Organization

### DISCLOSURE

The entire text references reporting requirements.

### APPLICATION SCOPE

Entity status: Private company , Public company , State-owned company  
Sectors:  
Description about target companies: This Standard can be used by an organization of any size, type, sector or geographic location that wants to report on its impacts related to this topic

### Tags

*Ilustración 38 - Ejemplo de detalles sobre una ley publicada y referenciada en la plataforma Datamaran. (Fuente: Datamaran)*

### Método para regulaciones obligatorias (leyes) (peso recomendado 30%)

El cálculo final del score que permiten clasificar los tópicos será:

$$V(t_j) = w_{ss} \cdot V_{ss} + w_{sf} \cdot V_{sf} + w_e \cdot V_e \quad (7)$$

$t_j$  es el tópico número  $j$  dentro del conjunto  $T$  incluyendo todos los tópicos definidos.

$V(t_j)$  es el volumen ponderado de  $t_j$ .

$w_{ss}, w_{sf}$  y  $w_e$  son los pesos de las regulaciones *sector specific*, *sector free* y *expected* respectivamente.

$V_{ss}, V_{sf}$  y  $V_e$  son los volúmenes absolutos (número de regulaciones en la base de datos) de los textos *sector specific*, *sector free* y *expected* respectivamente.

Una vez calculado el volumen ponderado de cada tópico, se calcula la media ponderada de los volúmenes correspondiente a los tópicos pertenecientes a la temática para calcular su volumen ponderado total, de acuerdo con la ecuación (5).

### Método para regulaciones voluntarias o recomendaciones legales (peso recomendado 20%)

Aquí se miden las recomendaciones legales, es decir textos legales que aplican incentivar (bonificaciones fiscales, reconocimiento oficial de un estatuto especial) al respetar una regla definida. El *score* correspondiente a esta fuente será simplemente el volumen absoluto de recomendaciones legales que entran en el enfoque definido en el análisis (véase Ilustración 37). En consecuencia, el *score* final de la temática será la media ponderada de los *scores* de cada uno de los tópicos que la componen, de acuerdo con la ecuación (5).

## 3.2.5. Noticias y redes sociales

Entre las diferentes fuentes que Datamaran usa para cuantificar la importancia externa de una temática, la opinión pública, medida por las redes sociales y los artículos de prensa ocupa una parte menor pero significativa.

La pestaña *Newsflow* a nivel global de la plataforma (véase Ilustración 11) permite introducir ciertos filtros y observar la evolución temporal del número de artículos relacionados con aquel enfoque.

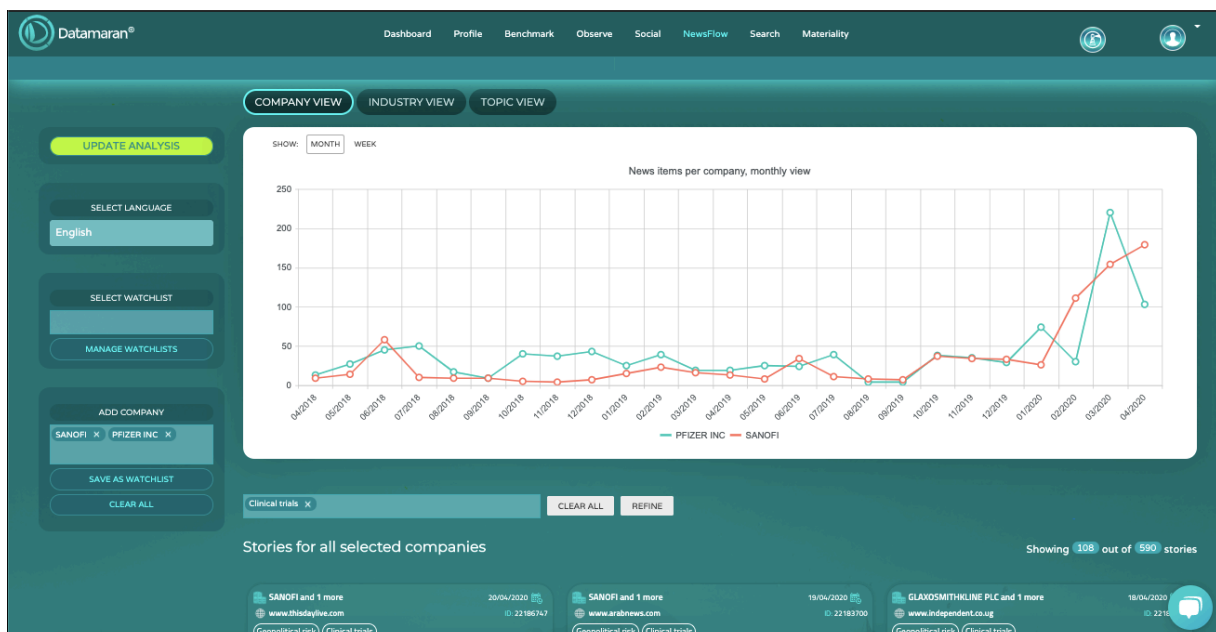


Ilustración 39 - Pestaña NewsFlow, vista por compañía. Gráfico: evolución del número de artículos mencionando SANOFI o PZIFER INC entre 04/2018 y 04/2020. (Fuente: Datamaran)

Tabla 5 - Repartición de los diferentes tipos de sectores entre las diferentes industrias (en negrita). (Fuente: Datamaran)

<b>Basic Materials</b>
Building Materials
Chemical Products
Mining and Metals
Packaging
Paper and Forest
<b>Consumer Goods and Services</b>
Accommodation, Catering and Leisure
Automobile Components and Manufacturing
Consumer products
Distribution
Domestic Durables
Food Retail
Food, Alcoholic and non-alcoholic beverages
Media, Publishing & Broadcasting
Miscellaneous Consumer Services
Retail
Tobacco
<b>Financial Services</b>
Banking Services
Capital Markets
Consumer Financials
Diversified Financials
Exchanges
Insurance
Other Financials
Real Estate Services
REITs
<b>Health Care and Pharmaceuticals</b>
Biotechnology, Life Sciences and Pharmaceuticals
Health Care Equipment and Services

<b>Industrials</b>
Air & Defense
Air, Rail and Road transportation
Building Products and Components
Business Support Services
Construction and Engineering
Electrical Equipment
Industrial Conglomerates
Industrial Machinery
Marine Transportation
Services and Supplies
Transportation Infrastructure
<b>Oil and Gas</b>
Oil and Gas and other Fossil Fuels
Oil and gas Equipment and Services
<b>Technology Equipment and Services</b>
Communications Equipment
Electronic Equipment and Components
Hardware
Internet Software and Services
IT Services
Semiconductors
Software
Telecommunications
<b>Utilities</b>
Alternative Power & Electricity
Electricity Utilities
Water, Gas and Diversified Utilities

La Tabla 5 presenta cómo se dividen las diferentes actividades comerciales de las empresas consideradas: primero entre industrias, y dentro las industrias existen subdivisiones llamadas sectores para mejor precisar la actividad de la empresa. La Ilustración 39 muestra primero las 3 diferentes vistas posibles de los artículos: Vista por compañía, vista por industria (véase Tabla 5), vista por tópico. Se ve cómo se puede seleccionar idiomas (solo inglés o castellano de momento), y elegir cuáles son las empresas que se quieren comparar en el gráfico. Debajo de la página se muestra un extracto de los artículos que sirvieron para calcular estas cifras, para que el usuario pueda leer el artículo ya que se dispone directo del artículo. Debajo del gráfico se ve un campo para filtrar estos artículos por tópicos, por ejemplo, aquí se han filtrado para obtener solo los que tratan del tópico ensayos clínicos.

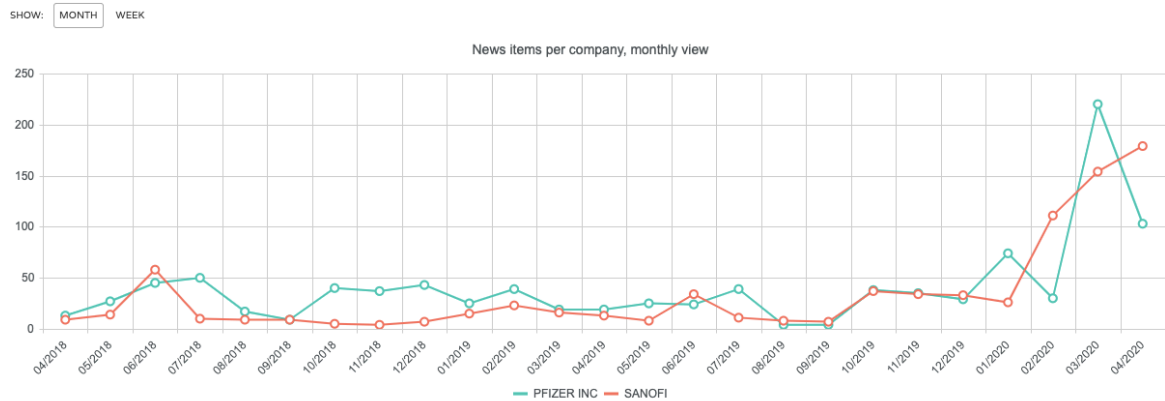


Ilustración 40 - Gráfico: evolución del número de artículos mencionando SANOFI o PZIFER INC entre 04/2018 y 04/2020. (Fuente: Datamaran)

El incremento que se observa en la Ilustración 40 al final corresponde a la pandemia de coronavirus que se observa claramente en el volumen de artículos hablando de las dos empresas farmacéuticas que se ha seleccionado: SANOFI (Francia) y PFIZER INC (Estados Unidos).

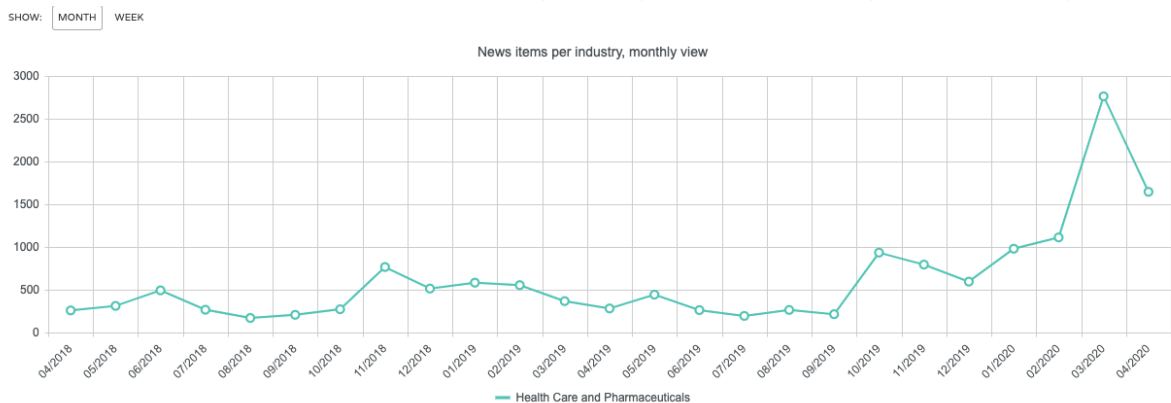


Ilustración 41 - Gráfico: Evolución temporal de las menciones de empresas pertenecientes a la industria de la salud y productos farmacéuticos entre 04/2018 y 04/2020. (Fuente: Datamaran)

Como lo muestra el gráfico de la Ilustración 41, la tendencia que se observó en la Ilustración 40 para ciertas empresas se generaliza a todo el sector de la salud. Este gráfico se puede obtener de la misma manera que se obtuvo el anterior, usando la pestaña de vista por industria, visible en la Ilustración 39.



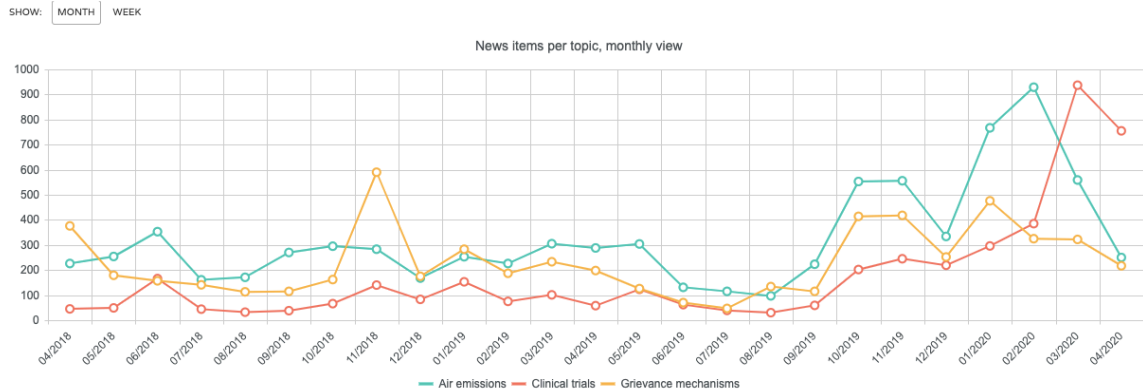


Ilustración 42 - Gráfico: Evolución temporal del número de artículos tratando de los tópicos: polución del aire, ensayos clínicos, mecanismos de reclamaciones. (Fuente: Datamaran)

En el gráfico de la Ilustración 42 se confirma, primero, la tendencia que se esperaba acerca de los ensayos clínicos: debido al coronavirus se incrementó significativamente el volumen de artículos acerca del tema. Los mecanismos de reclamación fueron también más importantes en octubre - noviembre 2019 ya que este tópico suele representar los momentos cuando el presidente Trump hace declaraciones polémicas. En este caso corresponde probablemente a sus declaraciones en reacción a la situación que provocó con el presidente de Ucrania (Molinares, 2019). Los artículos acerca de la polución del aire se relacionan por una parte con los altos niveles de polución detectados en febrero (Bruque, 2020), así como todas las discusiones sobre el posible impacto positivo del encerramiento debido a la pandemia sobre el cambio climático (León, 2020).

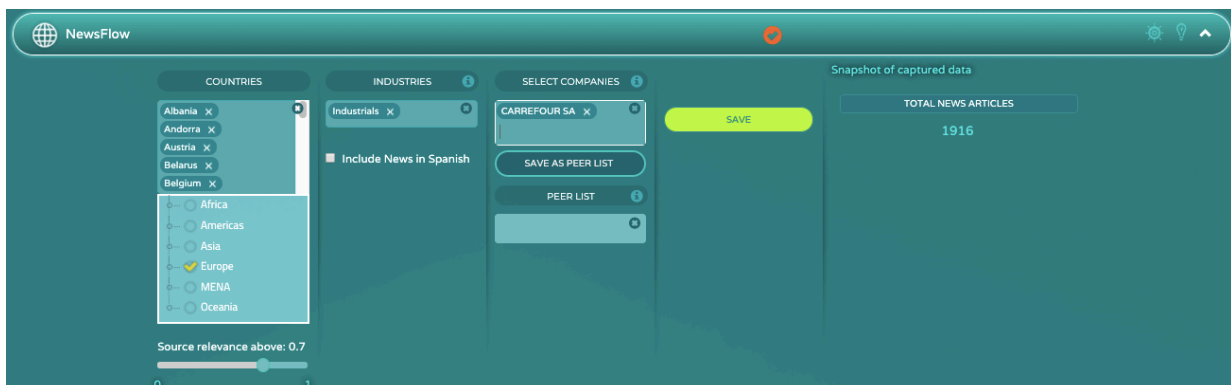


Ilustración 43 - Panel de parámetros de la fuente de noticias de prensa en la pestaña Materiality/External view. (Fuente: Datamaran)

La Ilustración 43 presenta los diferentes campos que se tienen que reseñar para parametrizar el análisis de noticias, se divide básicamente en 2 partes:

1. El filtro sobre las fuentes de noticias, incluyendo el campo *countries* y *source relevance*. Esta parte sirve para seleccionar solamente ciertos periódicos o páginas webs que producen noticias. Se puede hacer en función de su país de origen, o de la relevancia de aquella. La relevancia es una evaluación subjetiva realizada por Datamaran de los periódicos que proveen los artículos, será un número decimal entre 0 y 1, 1 siendo para los periódicos más fiables.
2. El filtro sobre las empresas consideradas en el artículo, incluyendo los campos *industries*, *select companies* y *peer list*. Estos campos no filtran las fuentes sino los artículos que tratan de ciertas compañías (que verifican al menos una condición del



filtro). Por ejemplo, en la Ilustración 43, se considerarán solamente los artículos que tratan de Carrefour, o de cualquier empresa teniendo una actividad industrial (véase la Tabla 5).

### Método de cálculo de score para artículos de prensa (peso recomendado 10%)

El *score* que corresponderá a esta fuente será simplemente el volumen absoluto acumulado durante 6 meses de artículos que entran en el enfoque definido en el análisis (véase Ilustración 43). En consecuencia, las variaciones de volumen puntuales, y de duración corta impactan poco el volumen total de artículos ya que se suma a 6 meses de datos. Finalmente, el *score* final de la temática será la media ponderada de los *scores* de cada uno de los tópicos que la componen, de acuerdo con la ecuación (5).

Acerca de la última fuente: redes sociales, la configuración es más sencilla en la versión actual de la plataforma ya que no existen filtros personalizables. Ciertos clientes como la BBVA pidieron por un filtro personalizado de Twitter con lo cual Datamaran es capaz de filtrar los tuits que hablan del sector específico de aquellos clientes, pero en general se usan todos los tuits a disposición para cuantificar la importancia de los tópicos.

La pestaña *Social* de la plataforma (véase Ilustración 11) sirve para testear y observar la evolución de las tendencias en la red social Twitter. No se puede enfocar a ciertas empresas, sectores de actividad ni países, pero permite comparar tópicos gracias a la funcionalidad de comparación de tópicos.

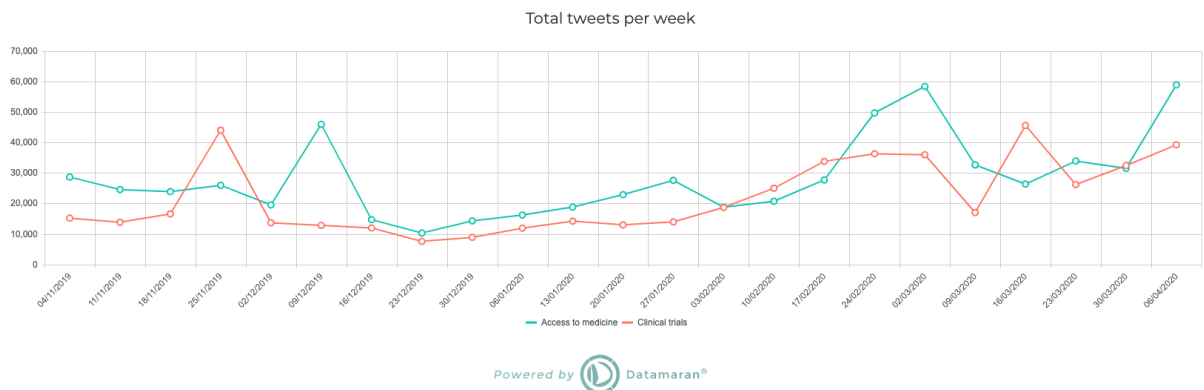


Ilustración 44 - Gráfico: evolución del número de tuits relacionados con los tópicos acceso a la medicina y ensayos clínicos. (Fuente: Datamaran)

La Ilustración 44 compara el número de nuevos tuits cada semana relacionados con los tópicos: acceso a la medicina y ensayos clínicos, que tienen comportamientos muy parecidos en general y también confirman la tendencia general provocada por la pandemia de 2020. Pero de manera general, comparar tópicos no aporta mucha información concreta, el usuario, en general, busca causas de los incrementos.



Ilustración 45 - Pestaña Social, vista detalles por tópico de los tuits acerca del tópico: acceso a la medicina - 20/04/2020. (Fuente: Datamaran)

La Ilustración 45 presenta la pestaña social, en la vista *Detalles por tópico* para el tópico *acceso a la medicina*, el primer gráfico (véase Ilustración 46) que presenta un desglose de los tuits entre el número de tuits originales únicos, y el número de contribuidores (retuits).

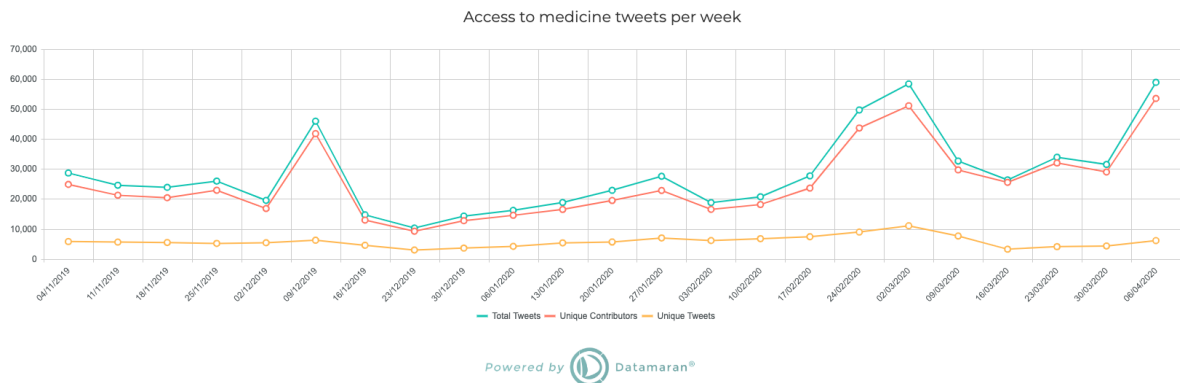


Ilustración 46 - Gráfico: evolución de los tuits semanales acerca del tópico: acceso a la medicina. - 20/04/2020. (Fuente: Datamaran)

Se puede observar en el gráfico de la Ilustración 46 que, si el número de tuits únicos no cambia mucho, el número de re-tuits o contribuidores pasa por periodos de picos donde pocos tuits se difunden a un gran número de personas.

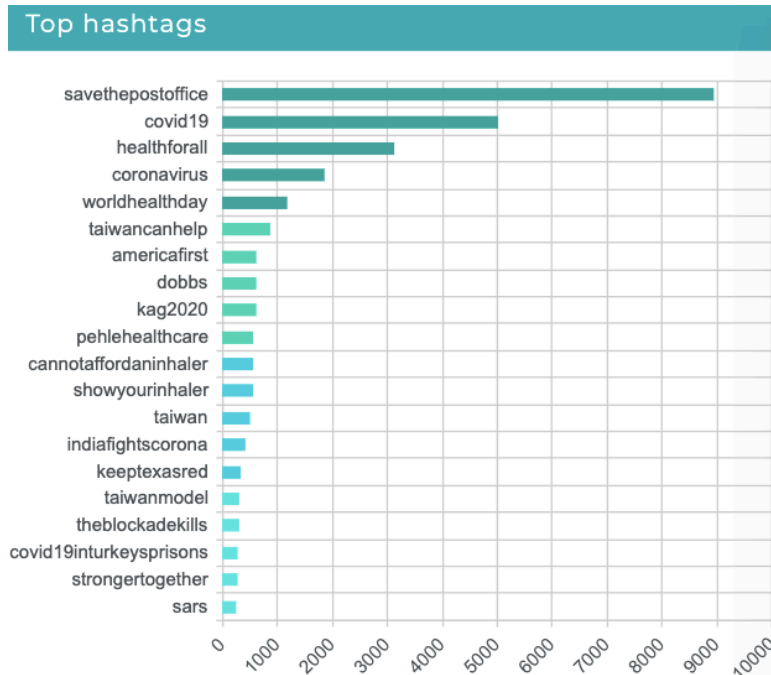


Ilustración 47 - Gráfico: hashtags más populares acerca del tópico: acceso a la medicina durante la semana del 06/04/2020. (Fuente: Datamaran)

El gráfico de la Ilustración 47 presenta los hashtags más populares del tópico, este gráfico probablemente será el más interesante para el usuario para entender rápidamente cuál es la causa de un alto número de tuits para una semana dada. En este caso, los *tops hashtags* tocan todos a la pandemia de Covid-19 y sus consecuencias sobre el sistema de correos en los estados unidos. La distribución de los volúmenes totales de tuits incluyendo estos hashtags es bastante interesante, como en el top 6 de los hashtags se observa un decremento exponencial del volumen conforme el *hashtag* ha sido menos compartido.

Debajo de estos gráficos se puede encontrar un extracto de algunos de los tuits más compartidos acerca del tema, y a la derecha de la página una columna con varias métricas, como una lista de palabras claves incluidas en la definición del tópico, el número total de impresiones<sup>13</sup> de todos los tuits.

Los tuits que se seleccionan solo provienen de fuentes que se validan por Datamaran manualmente, es decir, la empresa tiene una lista de vigilancia de cuentas verificadas (políticos, empresas o entidades oficiales) de las cuales se extraen en tiempo real todos los tuits.

#### Método de cálculo de score para redes sociales (peso recomendado 5%)

El *score* que corresponderá a esta fuente será simplemente el volumen absoluto acumulado durante 6 meses de tuits del tópico. En consecuencia, las variaciones de volumen puntuales, y de duración corta impactan poco el volumen total de tuits ya que se suma a 6 meses de datos. Finalmente, el *score* final de la temática será la media ponderada de los *scores* de cada uno de los tópicos que la componen, de acuerdo con la ecuación (5).

Las fuentes: artículos de prensa y redes sociales son las más inestables a corto plazo, es relevante observarlas semana por semana para ver cómo evoluciona la opinión en lo inmediato, pero por esas fluctuaciones, a medio y largo plazo estos movimientos generan ruido que no interesa al usuario.

<sup>13</sup> Este número corresponde al número total de veces que un tuit de la lista ha aparecido en la pantalla de un usuario de Twitter, haya interactuado con el tuit o no.

### 3.3. Continuous monitoring

Una vez el usuario tiene su análisis configurado y ejecutado, tiene acceso a su matriz y gráfico de materialidad (véase Ilustración 20 e Ilustración 22). más interesante que una fotografía del estado actual de la materialidad, un seguimiento temporal de ello aportaría mucha información útil para el cliente: ¿qué temáticas están ganando en importancia en mi entorno ahora? Para ello Datamaran 4<sup>14</sup> tiene un módulo especial que se llama el *continuous monitoring* y será el objeto de este trabajo.

#### 3.3.1. Caso de uso

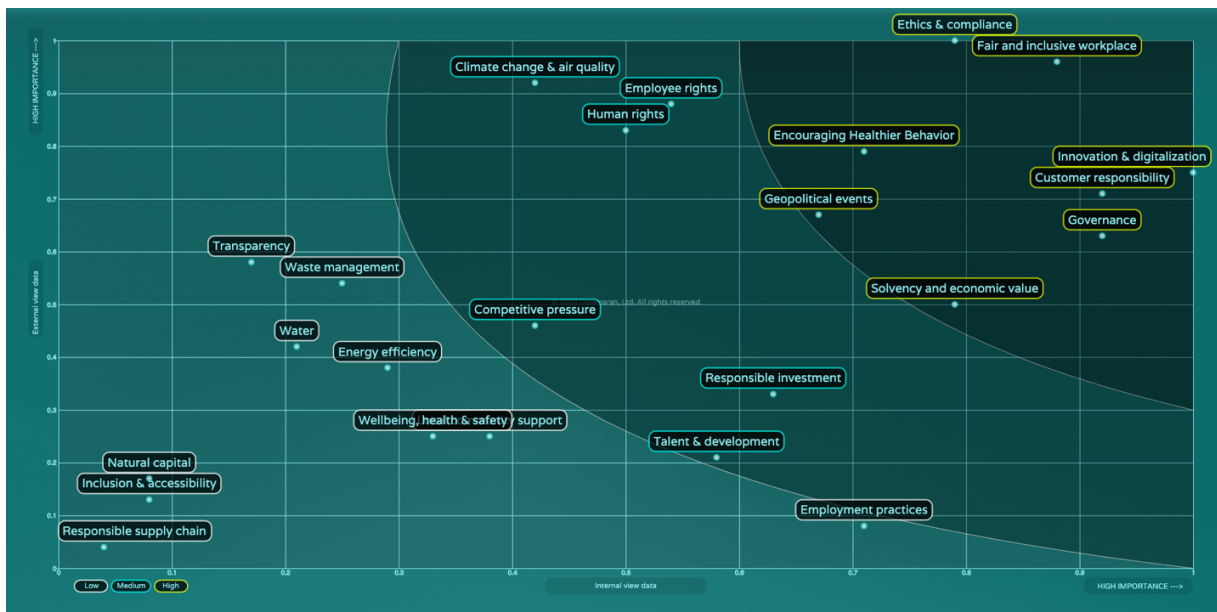


Ilustración 48 - Otro ejemplo de análisis de materialidad - vista gráfica. (Fuente: Datamaran)

El usuario tiene un análisis de materialidad como el de la Ilustración 48, se recuerda que el eje vertical representa la importancia externa (calculada a partir de los datos que Datamaran posee). El eje horizontal muestra la importancia interna a la empresa cliente de la temática considerada, parámetro que el cliente fija según sus criterios. El gráfico presenta la relevancia de cada temática, la combinación de esas dos importancias para sintetizar y clasificar las temáticas según el nivel de atención que se le debe de prestar. La variable que falta en este gráfico es el tiempo: no se puede ver dónde estaban las temáticas antes, y por tanto no se puede especular sobre dónde tienen más probabilidad de acabar el mes que viene.

<sup>14</sup> Versión de la plataforma a principios de 2020

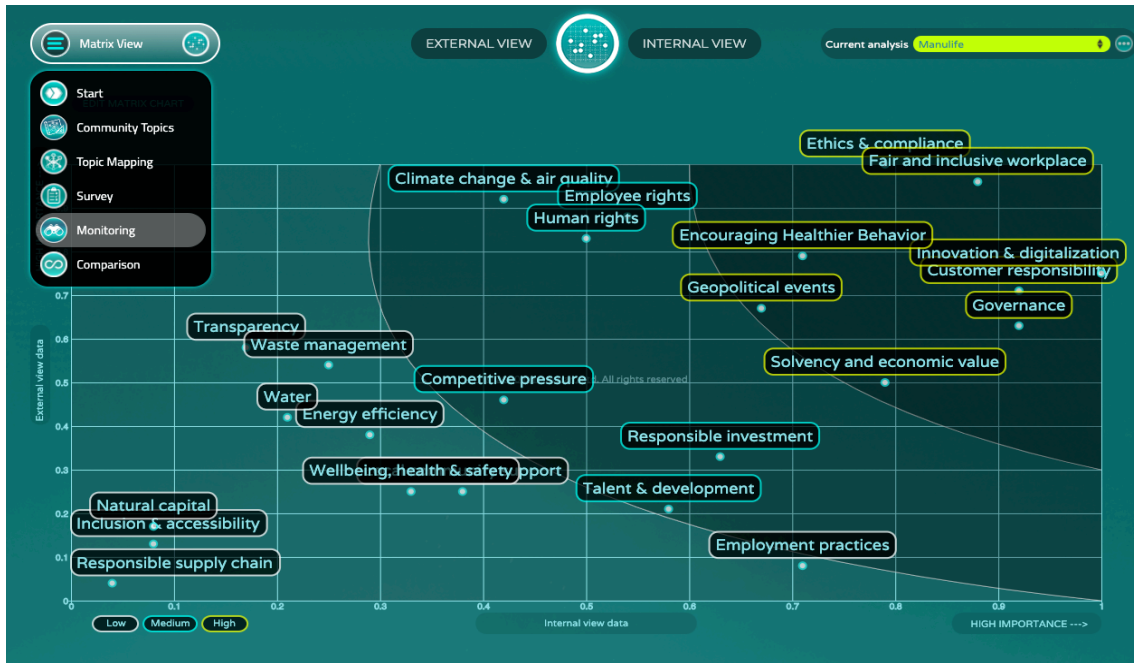


Ilustración 49 - Acceso al módulo de continuous monitoring desde la vista del análisis de materialidad. (Fuente: Datamaran)

Por tanto, Datamaran ya tiene un módulo , fácilmente accesible en la pestaña de materialidad cómo se ve en la Ilustración 49: *Continuous monitoring*, o *Monitoring*. El usuario podrá entonces acceder a un seguimiento temporal del estado de su análisis, tanto para constatar cuáles son las temáticas que han ganado más rangos últimamente, tanto como para saber cuales se tienen que vigilar. Una motivación importante para el cliente también es saber por qué dichas temáticas han incrementado su rango. Para ello, el módulo se descompone en dos partes, una lista de temáticas interesantes bajo forma de tabla, y un gráfico interactivo, cómo se puede observar en la Ilustración 50 a continuación.

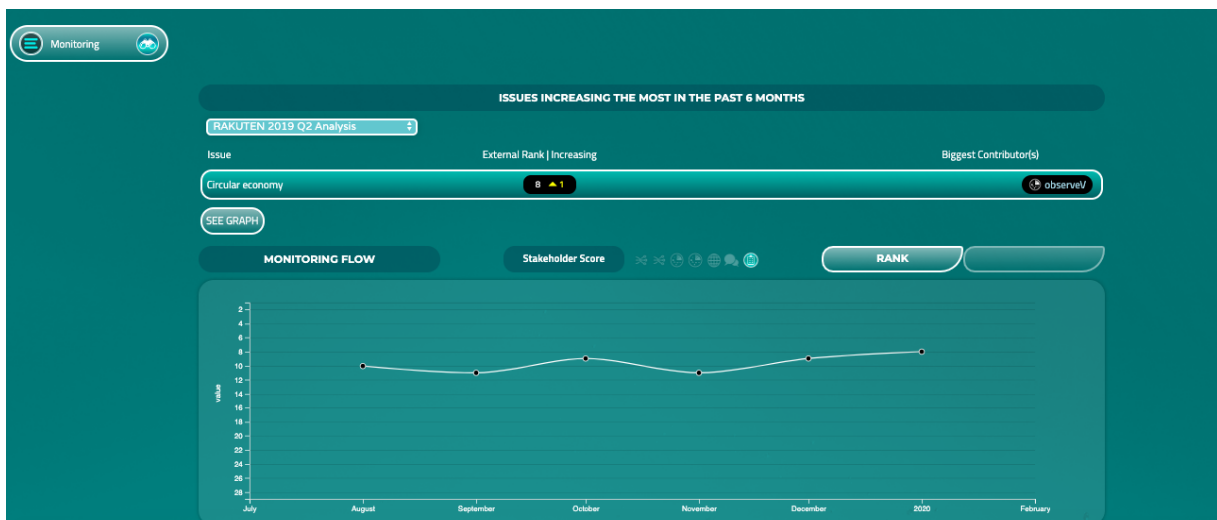


Ilustración 50 - Ejemplo de vista de continuous monitoring con solo una temática incrementando. (Fuente: Datamaran)

En el caso preciso del análisis de la Ilustración 48, el usuario puede obtener una vista de la tabla de *continuous monitoring* como la de la Ilustración 51. Esta tabla presenta las 5 temáticas que incrementaron más su rango en los últimos 6 meses. También se presentan los contribuidores

de este incremento, será una lista de las fuentes que usa el usuario. En este caso, por ejemplo, la temática *Employment practices* después de haber ganado 5 rangos, se encuentra en la posición 8. (el usuario ha definido 24 temáticas en total). Este incremento se explica sobre todo por un incremento de la importancia de la temática en los artículos de prensa y en los informes de sus competidores (o empresas de referencia elegidas por el usuario).

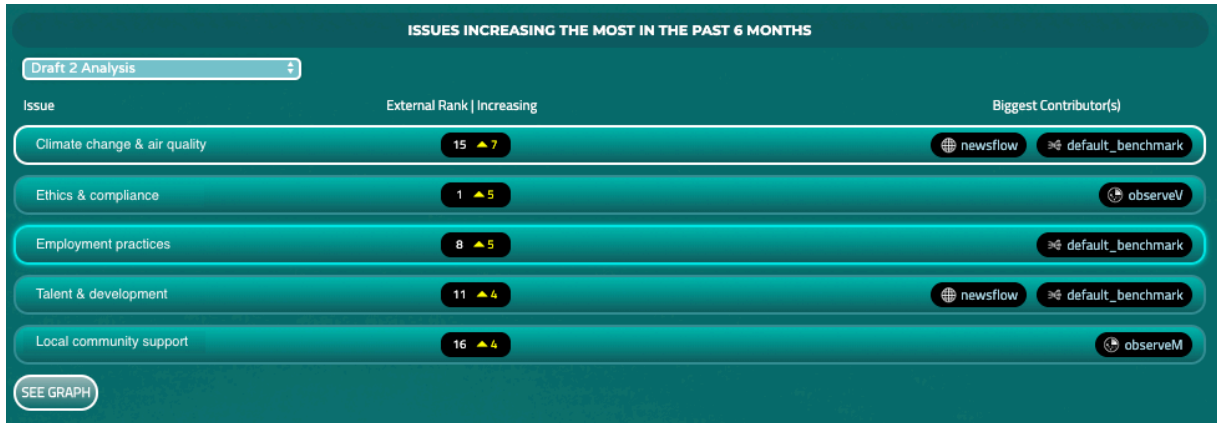


Ilustración 51 - Lista/Tabla de Continuous monitoring que presenta las 5 temáticas que incrementaron más en rango. (Fuente: Datamaran)

Si el usuario desea observar cómo ha evolucionado esta temática en los últimos 6 meses se puede abrir el gráfico con el botón *See graph* y ver la representación de la Ilustración 52. El gráfico presenta la posición del *stakeholder score* (o *score* total externo de la temática), por eso el eje vertical está invertido: cuanto más cerca de 1 mejor.



Ilustración 52 - Evolución del score externo de la temática *Employment practices* para el análisis de la Ilustración 43. (Fuente: Datamaran)

El gráfico también propone estudiar las evoluciones de cada una de las fuentes, tanto en su posición, como en score bruto para intentar entender mejor cómo se pueden explicar los incrementos.



### 3.3.2. Arquitectura y procesos involucrados

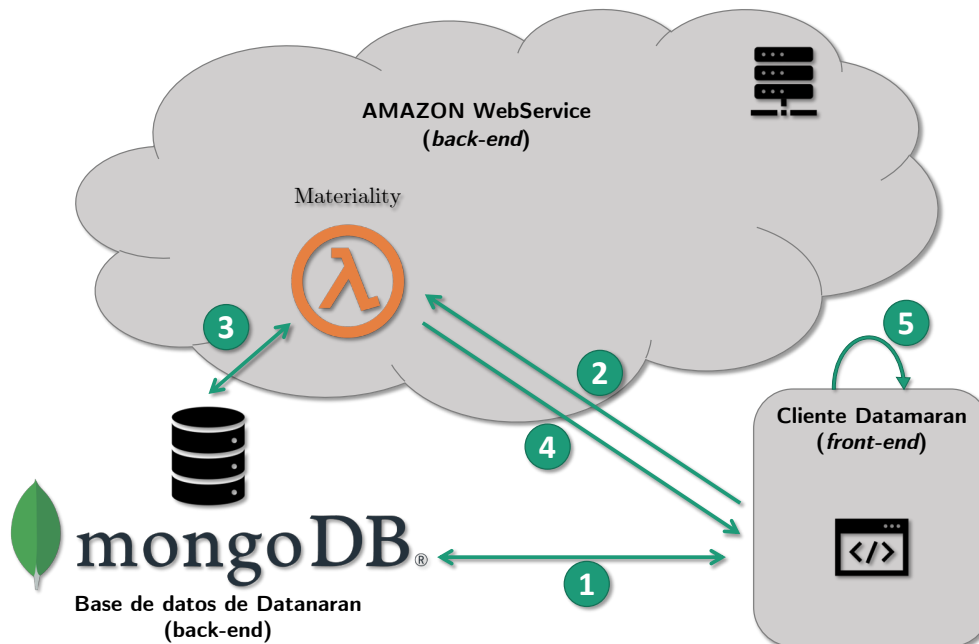


Ilustración 53 - Esquema de comunicaciones entre las diferentes partes del software Datamaran. (Fuente: elaboración propia).

La Ilustración 53 representa un esquema simple de como funciona el modulo de *continuous monitoring* actualmente. Desde el lado cliente (*front-end*), se recupera los datos de un análisis seleccionado por el usuario en la base de datos. La descripción del análisis, del enfoque de cada fuente luego se usa para llamar a la *Lambda* de Amazon WebService, ubicada en el lado servidor (el *cloud* aquí). La *Lambda* enviara pedidos a la base de datos y recopilara los resultados para cada tópico para los 6 meses anteriores. Una vez acabado, esto se envía al lado cliente, que hará las calculaciones, necesarias para enseñar la evolución de las temáticas, los puntos en el grafico y los contribuidores principales. Finalmente, este mismo lado enseñara los resultados al usuario de manera mas visual (usando los diseños del equipo de *Webdesign*).

## 3.4. Problemas y oportunidades encontrados

La herramienta de *continuous monitoring* proporciona una visión interesante para el usuario, pero falta de robustez, profundidad y perspectiva.

### Robustez

En el mundo del desarrollo web, una buena práctica es separar el código *front-end* (la experiencia de usuario) del *back-end* la información que se muestra. Esta práctica se justifica sobre todo por razones de escalabilidad de la producción, para que el código *front-end* (cargado en la página) sea lo más ligero posible. En su estado actual, el módulo está integralmente escrita en *front-end* – a parte de los cálculos de materialidad que se hacen en *back-end* el cálculo de los puntos de gráfico, de incrementos de rango y de contribuidores se hacen en directo, lo que puede ralentizar el funcionamiento de la aplicación.

Cómo se ve en la Ilustración 52, este mismo código tiene problemas para mostrar el mes de enero y muestra el año 2020, eso ha impactado la lógica de cálculo y ha provocado bugs en principios de año por ello.



### Profundidad

El cálculo de contribuidores aporta una buena información al usuario sobre dónde se sitúan las causas de un incremento de importancia, pero no es suficiente. El objetivo último de la búsqueda de causas es encontrar los artículos, leyes, o informes que fueron difundidos últimamente y provocaron las variaciones.

Los contribuidores se calculan considerando lo siguiente:

$$\text{Máxima contribución}(T_i) = \max_{f \in \text{Fuentes}} (w_f \cdot [Pc(T_i, f, t) - Pc(T_i, f, t_0)]) \quad (8)$$

$T_i$  es la temática número  $i$ .

$f$  es una fuente dentro del conjunto *Fuentes*.

$t_0$  es el momento inicial del estudio,  $t$  el momento final.

$w_f$  es el peso atribuido a la fuente  $f$ .

$Pc(T_i, f)$  es el *percentile rank* de  $T_i$  en el momento  $t$ .

$w_f \cdot (Pc(T_i, f, t) - Pc(T_i, f, t_0))$  es la contribución de la fuente  $f$  para la temática  $T_i$ .

El (o los) contribuidores más importantes serán las fuentes  $f$  que tienen una contribución igual a la máxima contribución definida en la ecuación (8). Este cálculo tiene sentido para entender qué fuente contribuye más para incrementar el *stakeholder score*, pero después de clasificar las temáticas, este concepto de score es relativo. Una temática  $T_1$  con un score dado cierto mes tendrá una posición  $N$  este mes, el mes siguiente, si una temática  $T_2$  disminuye su *score* suficientemente, esta última puede acabar en la posición  $N$  y entonces hacer que  $T_1$  suba a la posición  $N - 1$  (mejor que  $N$ ). Por esta razón es posible observar que algunas temáticas ganan posiciones sin que ningún contribuidor pueda ser identificado (y en este caso se enseña un mensaje diciéndole).

En conclusión, desde el punto de vista del usuario sería interesante tener más indicaciones sobre las razones que hacen que una temática cambie de posición.

### Perspectiva

Si la herramienta que ayuda a ver cómo han evolucionado las temáticas interesa a los clientes, es sobre todo para realizar previsiones. Las observaciones del pasado, relacionadas con consecuencias reales en la vida de la empresa (pérdida de clientes, escándalos) se pueden usar para mejorar su estrategia frente a los riesgos. Gracias a un diagnóstico constante de la situación, y un panel de indicadores se puede esperar obtener más visión de lo que el futuro anuncia acerca de las temáticas. De momento, el módulo tal como es no propone nada de este estilo, aunque sería algo que interesa muy fuertemente a los clientes.

En este trabajo se tratará de resolver primero el problema de robustez con una nueva arquitectura y un nuevo diseño del módulo, para que sea más atractivo y también más organizado. También se tratará de ofrecer más perspectiva ofreciendo una visión diferente sobre los datos actuales y tratar de interpretarlos para mejorar preparar a los clientes al futuro. Desafortunadamente, el problema de profundidad de la herramienta no se tratará en el presente trabajo, pero constituirá una de las líneas futuras de trabajo sobre este módulo.

## 3.5. Conclusión

La herramienta de Datamaran es sirve principalmente para generar un análisis de materialidad del entorno de la empresa del cliente. Para ello es necesario definir y clasificar las temáticas éticas consideradas tanto desde el punto de vista del cliente como de su entorno exterior. El punto de vista interno se define solamente por el usuario según sus criterios, pero el externo se tiene que definir a través de las diferentes fuentes accesibles. Cada una de esas fuentes tiene una pestaña



dedicada que permite testear los filtros que se quieren aplicar de cara a definir el enfoque del análisis. Una vez el usuario tiene su análisis parametrizado, se recuperan los datos de las bases de datos de Datamaran y se calculan los datos que servirán a clasificar en fin las temáticas.

Si esto permite tener una idea clara de la posición de las temáticas y los riesgos asociados, hace falta una visión diferente para poder seguir su evolución en el tiempo. En la versión actual de la plataforma existe un módulo llamado *Continuous monitoring* que realiza este seguimiento. Aquel módulo ha sido implementado rápidamente en respuesta a una necesidad obvia por parte de los clientes, pero nunca se ha construido algo robusto que facilite las explicaciones y diagnósticos profundos que se esperan de ello.

El presente trabajo tratará de resolver los problemas asociados con la robustez y perspectiva del módulo gracias a un remodelado de su funcionamiento y añadiendo las herramientas necesarias para aportar valor a los usuarios.



## 4. Definición de los requerimientos del módulo de monitoring

### 4.1. Introducción

Se ha definido como funcionaba la plataforma de Datamaran y más especialmente el proceso de cálculo que permite obtener el análisis de materialidad. Este análisis constituye la mayor parte del valor para los usuarios, pero el seguimiento de la evolución de los datos y las causas de su evolución interesa a los clientes.

Este capítulo tratará de definir los requerimientos del proyecto, aquellos requerimientos vienen por la mayoría de otros departamentos de Datamaran, que se encargan de traducir las necesidades de los usuarios/clientes. A partir de estos requerimientos se empezará a diseñar una solución para el módulo de *monitoring*.

El capítulo presenta primero a los actores del proyecto para precisar los diferentes roles. Luego los requerimientos se dividirán entre, por un lado, los que se exigen por que no implican mucha creatividad en el diseño de la solución y, por otro lado, se describirán los objetivos abiertos a más largo plazo que se esperan alcanzar con el remodelado de la herramienta de *continuous monitoring* para Datamaran.

### 4.2. Actores del proyecto

Los clientes internos del proyecto son miembros del equipo responsable del producto y del entrenamiento de los clientes al uso del *software*. Tienen una visión de la plataforma, más específicamente los flujos de experiencia de usuario, o cómo se debe desarrollar las distintas etapas del uso de la plataforma por un usuario. Por tanto, tienen ideas sobre cómo se deben poder usar los módulos y sobre lo que tiene que aparecer al usuario.

Tabla 6 - Lista de actores del proyecto y sus roles en Datamaran. (Fuente: elaboración propia)

Tipo de actor	Rol
Product Owner	Director of Customer Success
	Director of Product
	CTO
Scrum Master	Main Software Architect
Responsables	Senior Data Scientist
	Junior Data Scientist
Miembros del equipo	Developers
	Quality Assurance
	Web design/UX

La Tabla 6 presenta los diferentes actores, los *Product Owners* serán los que se considerarán como los clientes, los que definen y priorizan las funcionalidades que añaden más valor. El *Scrum Master* será el facilitador del proyecto, que tiene una visión global de lo que se tiene que hacer y ayuda al equipo. Los otros actores serán, por un lado, los responsables, que propondrán las

soluciones a los *Product Owners*, y, por otro lado, los miembros del equipo, los que efectúan las tareas durante los *sprints*.

## 4.3. Requerimientos obligatorios

### 4.3.1. Requerimientos funcionales

En este apartado se van a presentar todos los requerimientos en el funcionamiento y diseño de la página, desde el punto de vista del usuario. En la metodología *Scrum* (véase 1.2.2) de desarrollo de los proyectos, se suele pensar poniendo el usuario en el centro de las preocupaciones. Por tanto, siempre se describen los requerimientos bajo la forma de *user stories*, o experiencias de usuario, que tienen una estructura bien precisa: *as a [type of stakeholder], I want to...* En este caso el único stakeholder que usa la página es el usuario (no hay administrador de cuentas, ni responsable SEO<sup>15</sup>). Entonces una *user story* en español sería: como usuario, quiero...

Tabla 7 – Requerimientos para la tabla de monitoring, presentados por el equipo de producto que se tienen que realizar obligatoriamente en la nueva versión del módulo . (Fuente: elaboración propia).

Nº del requerimiento	Tipo de requerimiento	Requerimiento (user story)
R1	Display	Como usuario, quiero ver cuál es la relevancia de las temáticas que estoy observando.
R2	Display	Como usuario, quiero poder ver la situación de todas las temáticas
R3	Data	Como usuario quiero poder ver las fuentes responsables del cambio
R4	Data/display	Como usuario, quiero poder comparar la situación actual con los datos de mi análisis en su estado en el momento del último cálculo
R5	Data	Como usuario, quiero poder descargar un fichero Excel conteniendo los datos de la página
R6	Display	Como usuario, quiero poder seleccionar simplemente cualquiera de mis análisis sin cambiar de página.
R7	Data/display	Como usuario quiero ver la posición actual de cada temática y de cuanto ha incrementado.

La ventaja principal de las *user stories* es que resumen una funcionalidad entera que la página debe de tener, y no una etapa de desarrollo: obliga al desarrollador a pensar como un usuario y publicar solo trozos de códigos que funcionan independientemente de que la página esté acabada. Sería como decir que una casa se tiene que realizar pieza por pieza: primero el baño, luego la

<sup>15</sup> *Search Engine Optimization*: generalmente se encargan de redactar artículos, añadir palabras claves y descripciones para mejorar la visibilidad de una página dentro de los resultados de búsqueda en *Google*, por ejemplo.

cocina, pero obviamente, este método se puede aplicar en el mundo informático, y no en el mundo de la construcción.

La Tabla 7 presenta una lista de todos estas *user stories* que se tienen que realizar para la parte alta de la página (tabla de temáticas), numeradas y clasificadas según el tipo de trabajo que necesitan:

- Display significa que la mayor parte del trabajo está en las interacciones visuales, no se ejecutan cálculos, solo se desplaza y se enseñan los datos. Por tanto, será responsable de su realización el equipo *front-end*.
- Data significa que hay un trabajo importante de cálculo y de manipulación de los datos que se va a realizar fuera del entorno de la página. Por tanto, será responsable el equipo *back-end*.
- Data/display significa que ambos serán importantes, y que la comunicación será importante entre los dos. En este caso se considerará responsable el equipo *back-end*, y se tendrá que encargar de la buena coordinación de ambos equipos.

\*Variation in ranking for this issue is only due to changes in other issues

Ilustración 54 - Esquema (mockup) de la nueva disposición de la tabla de temáticas. (Fuente: elaboración propia).

En la Ilustración 54, que se encontrará en la parte alta de la página, donde la antigua tabla (véase Ilustración 51) presenta las diferentes temáticas, su posición hoy en día y su evolución se quiere aportar ciertos cambios.

El requerimiento R1 corresponde a la relevancia de la temática, es decir la zona de importancia que le corresponde en el gráfico representando el análisis de materialidad de la Ilustración 20.

El requerimiento R2 tiene que cumplirse gracias a un botón que permitirá extender la tabla de temáticas. Por defecto solo se ven el top 5 de temáticas que han ganado más posiciones durante los últimos 6 meses.

Para entender el requerimiento R3, y porque es diferente de lo que ya existía, hay que recordar que, en la versión actual del módulo, solo se enseña el top 5 de las temáticas que mejoraron su posición. Si tal vez, el análisis no tiene 5 temáticas que incrementaron en posición, solo se enseñan aquellas que incrementaron. Con lo cual nunca se enseñaría una temática estática o que empeora, por tanto, nunca se ha tratado de calcular cuál es el mayor contribuidor de aquellas. Entonces el requerimiento R3 incluye realmente 3 aspectos:

1. Si la temática mejoró, se trata de reproducir como mínimo lo que existe actualmente

- para los contribuidores
2. Si la temática es estática (no modificó su posición), no se debe mostrar nada en la columna *drivers of change*
  3. Si la temática bajo su posición se debe mostrar las fuentes responsables del decremento (no será la misma lógica que para un contribuidor).

En ciertos casos, es posible que no encuentre ningún *driver* del cambio, con lo cual se tendrá que señalar con un asterisco y el texto escrito para explicar que los cambios se produjeron debido a los cambios de las otras temáticas.

El botón de tipo *toggle* del requerimiento R4 debe permitir cambiar la visión de toda la página:

- En posición *6 months view* debe enseñar la comparación entre la situación actual y como era hace 6 meses. Eso es exactamente lo mismo que en la versión antigua del módulo .
- En posición *Last refresh view* debe comparar los datos del análisis en el momento de su última ejecución con la situación calculada hoy en día. En su pestaña de materialidad el usuario tiene un análisis calculado con ciertos parámetros y estos resultados no cambian con el tiempo. Entonces con este punto de vista los usuarios podrán comparar las cifras que tienen actualmente en su análisis de materialidad con los datos más actuales que tiene Datamaran, usando los mismos parámetros de enfoque.

En la Tabla 7, el requerimiento R5 representa la necesidad de tener un botón para descargar un archivo Excel conteniendo todos los datos que se presentan en la tabla (para ambas situaciones: *6 months view* y *last refresh view*).

Para cumplir el requerimiento R6 solo se necesita una lista desplegable conteniendo una referencia a todos los análisis presentes en la cuenta del usuario. Por fin el requerimiento R7 es básicamente la función principal de la herramienta, tal y cómo se podía ver en su versión antigua. Se tiene que ver la posición externa actual de la temática, así como la cantidad de posiciones incrementadas durante el periodo de tiempo seleccionado.

Tabla 8 - Requerimientos para el gráfico de continuous monitoring. (Fuente: elaboración propia)

Nº del requerimiento	Tipo de requerimiento	Requerimiento (user story)
R8	Data/display	Como usuario, quiero ver fácilmente en el gráfico la evolución entre 2 puntos del gráfico cuando los selecciono
R9	Display	Como usuario, cuando selecciono una temática, quiero ver el gráfico que corresponde con una de las fuentes responsables por el cambio.
R10	Display	Como usuario, quiero ver un gráfico que representa la evolución de cada fuente, tanto en posición como en volumen bruto de datos

La Tabla 8 presenta más específicamente los requerimientos relacionados con la parte baja de la herramienta: el gráfico. En la *mockup* de la Ilustración 57 se pueden apreciar las ubicaciones de cada uno de ellos, incluso el requerimiento R4 que era parte de la vista anterior. Aquel requerimiento hace referencia aquí a la línea vertical amarilla que tendrá que marcar la fecha de última ejecución del análisis de materialidad. No obstante, el gráfico permanecerá igual en las dos posibles vistas (véase botón del requerimiento R4 en la Ilustración 54), porque el gráfico representa toda la evolución temporal, no solo una comparación de 2 puntos.



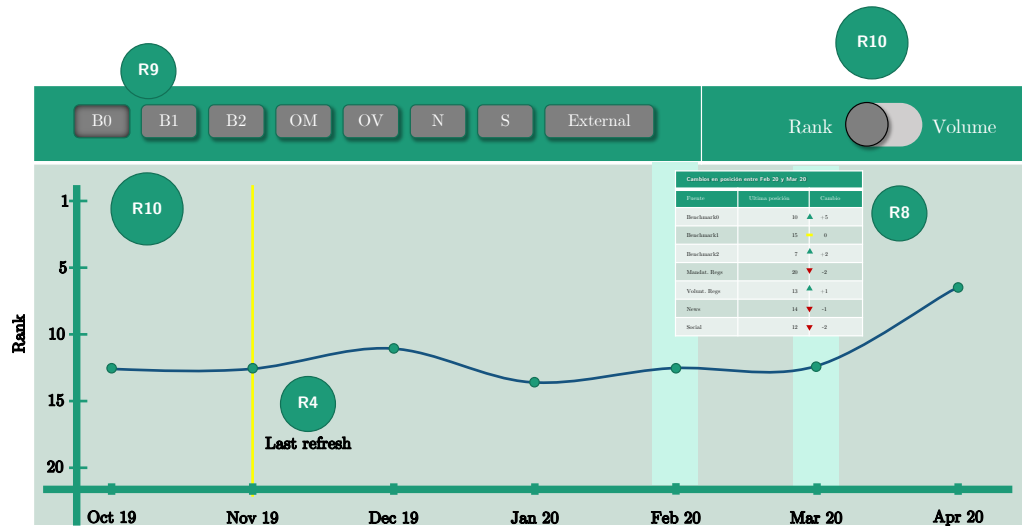


Ilustración 55 - Esquema (mockup) de la situación deseada para la parte baja de la página, el gráfico. (Fuente: elaboración propia)

El requerimiento R8 significa que cuando un usuario selecciona un punto del gráfico, primero se debe marcar con fondo diferente (véase las columnas claras de la Ilustración 55). Segundo, cuando otro punto se ha seleccionado debe aparecer una pequeña tabla de indicaciones o *tooltip*<sup>16</sup> conteniendo los datos de evolución de todas las fuentes.

Tabla 9- Tooltip que debe aparecer al seleccionar dos puntos distintos en el gráfico de monitoring. (Fuente: elaboración propia).

Cambios en posición entre Feb 20 y Mar 20		
Fuente	Última posición	Cambio
Benchmark0	10	▲ +5
Benchmark1	15	■ 0
Benchmark2	7	▲ +2
Mandat. Regs	20	▼ -2
Volunt. Regs	13	▲ +1
News	14	▼ -1
Social	12	▼ -2

Como lo muestra la Tabla 9, el requerimiento R8 pide que para cada posible gráfico y cada pareja de puntos se enseñen los cambios para todas las otras fuentes. Es importante mencionar que los cambios se tienen que calcular siempre en el sentido cronológico naturalmente, independientemente del orden de selección del usuario.

<sup>16</sup> Ayuda visual que se activa en general cuando se pasa el ratón encima

Tabla 10 - Correspondencia de nombre y códigos para el requerimiento R9. (Fuente: elaboración propia).

Nombre de la fuente	Código usado en la Ilustración 55
Benchmark 0	B0
Benchmark 1	B1
Benchmark 2	B2
Mandatory Regulations	OM
Voluntary Regulations	OV
Newsflow	N
Social	S
<b>External score/Stakeholder score</b>	<b>External</b>

El requerimiento R9 obliga a tener una lista de botones representando todas las fuentes (con su correspondiente código en la Tabla 10). Un clic al botón de la fuente debe directamente enseñar el gráfico correspondiente. Otra parte del requerimiento es que al seleccionar una temática en la tabla arriba debe de enseñar automáticamente un gráfico correspondiente con una de las fuentes responsables por el cambio. Entre todas las fuentes también aparece la que se llama *External* o *stakeholder score*, que representa una combinación de las posiciones de todas las otras fuentes ponderadas por el peso que se le atribuyó el usuario.

Por fin el requerimiento R10 exprime la necesidad de ver el gráfico adaptado para cada pareja temática – fuente, y dejar la posibilidad al usuario, a través de 2 botones de cambiar de la visión absoluta (*Volume*) a la visión relativa (*Rank*). Una diferencia notable me pareció importante destacar es el número de puntos en el gráfico comparado con el número que había en la primera versión del módulo . Observando la Ilustración 52 se nota que aparecen 6 puntos, correspondiente a los 6 meses de vigilancia que se pretenden ofrecer al usuario. Tras pensarlo cierto tiempo propuse una corrección a esto, dado que cuando se observan 6 puntos, solo se observan 5 intervalos. La diferencia de posición entre el primero y el último mes representan solamente 5 meses de incremento. Por tanto, se ha definido en los requerimientos de la nueva versión que se estudiarán 7 puntos, correspondiente a 6 intervalos de un mes cada uno.

### 4.3.2. Requerimientos no funcionales

Para construir una aplicación de tamaño relativamente grande es importante que la arquitectura quede bien estructurada y escalable. Eso implica entre otros separar el diseño y los que trabajan sobre ello en 3 partes distintas:

1. Los *web-designers* & *UX designers* se encargarán de los visuales (logos, colores, estilo global de la página y sus módulos)
2. Los desarrolladores *front-end* se encargan de escribir el código que anima la página y sirve de interfaz gráfica con el usuario. Son la parte dinámica que da a la plataforma un aspecto interactivo.
3. Los desarrolladores *back-end* se encargan de facilitar los datos correctos al *front-end*, ejecutan los cálculos de forma escondida, en general del lado del servidor para que sea más rápido

Este requerimiento no viene de ningún cliente, tanto externo (usuarios en las empresas) como internos (miembros responsables del producto, ventas, marketing), sino que viene del propio departamento técnico. En efecto, una buena arquitectura en general no crea valor directamente para el usuario, pero reduce significativamente los costes en tiempo de mantenimiento, *debug* y los riesgos de averías.

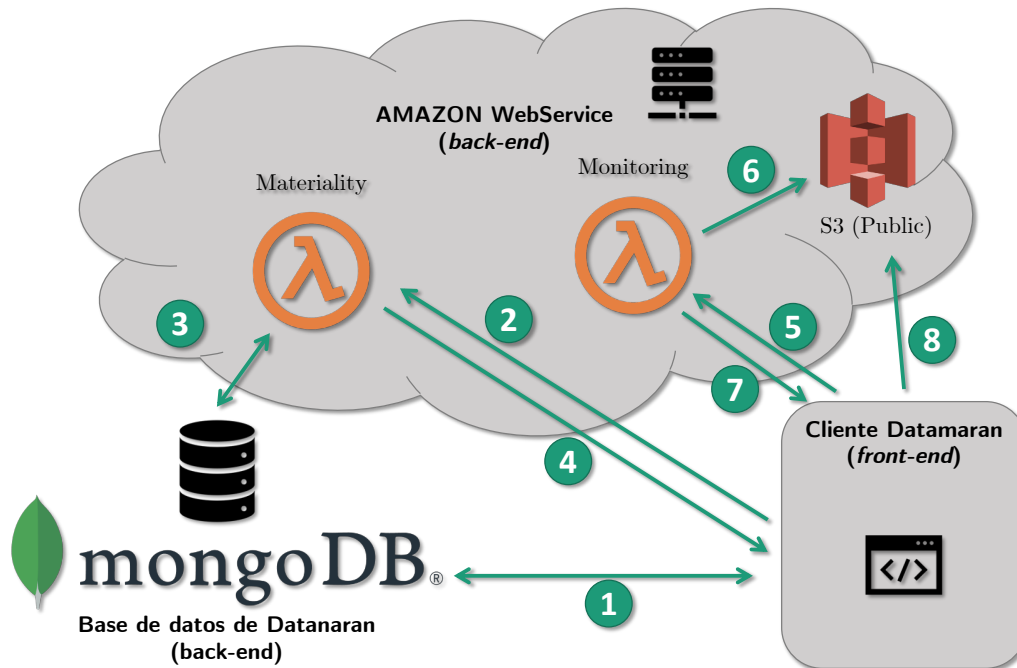


Ilustración 56 - Esquema de comunicaciones entre las diferentes partes del software Datamaran. (Fuente: elaboración propia)

En la Ilustración 56 se presenta la nueva arquitectura. Se diferencia del esquema de la Ilustración 53 sobre todo por la creación de una nueva función *lambda* que será la de monitoring. El departamento técnico requiere esta arquitectura para poder segmentar lo más posible las actividades – separar los cálculos de la experiencia de usuario.

En esta nueva arquitectura el proceso será el siguiente:

1. El *front-end* llama la base de datos para recuperar los datos correspondientes del análisis. Aquí el requerimiento R6 tiene sentido porque requiere llamar la base de datos cada vez que se cambia de análisis. Los datos que se recogen son básicamente la definición del enfoque de cada una de las fuentes para poder calcular el análisis de materialidad.
2. Los parámetros recuperados en 1 se envían a la *lambda* de materialidad, que los recibe, corrige e interpreta. La diferencia con la antigua arquitectura también está en que en la etapa 2 se piden directamente los datos para todos los meses que se necesitan. Antes se pedían los 2 extremos (más reciente y antiguo) para calcular la tabla, y luego cuando se pedía el gráfico se calculaban los puntos restantes.
3. Se pide a la base de datos todos los elementos que entran en el enfoque del análisis y se realizan los cálculos que permiten evaluar cada temática, eso para cada uno de los 7 puntos que se han pedido.
4. La *lambda* responde al *front-end* con los resultados de *scores* para cada tópico
5. El *front-end* reorganiza los datos recibidos en 4, agrupa los tópicos en sus temáticas y envía a la *lambda* de *monitoring*
6. Se efectúan los cálculos en el *cloud*, y cómo se espera un archivo Excel que descargar, este se crea y se envía al servicio S3 de AWS<sup>17</sup>. Este servicio permite almacenar archivos temporalmente en el *cloud* que se podrán descargar posteriormente desde una página web remota.
7. Se responde al *front-end* con los datos calculados y organizados, más un camino hacia el archivo que se almacenó en S3.

<sup>17</sup> Amazon WebService

8. El vínculo hacia el archivo almacenado en S3 permite descargar el archivo desde el *front-end*.

## 4.4. Requerimientos libres

El equipo de entrenamientos ha creado una presentación bajo la forma de un PowerPoint personalizable, que se facilita a los clientes. Esta plantilla se destina a los responsables de la política de sostenibilidad dentro de las empresas clientes para que comuniquen con su dirección los aspectos relevantes del entorno no financiero actual. El objetivo de una presentación de este tipo es complementar la plataforma con un estudio más profundo de los datos presentados. Idealmente, este análisis se debería poder hacer automáticamente y extraer varios *insights* de los datos que se manejan en Datamaran.

Tabla 11 - *Requerimientos abiertos para el nuevo módulo de monitoring, acerca del diagnóstico automático de los análisis. (Fuente: elaboración propia)*

Nº del requerimiento	Tipo de requerimiento	Requerimiento (user story)
R11	Data	Como usuario quiero poder identificar las temáticas que están de moda hoy en día.
R12	Data	Como usuario quiero poder identificar las señales sociales que pueden demostrar de una debilidad de mi empresa en ciertas temáticas.
R13	Data/display	Como usuario quiero obtener información sobre los mecanismos que provocan un incremento de mis temáticas en importancia:

La Tabla 11 presenta los requerimientos del módulo que no tienen una solución bien definida, son más abiertos y se deja la libertad de definir la metodología para cada uno.

Una de las diapositivas de la plantilla PowerPoint de Datamaran se presenta en la Ilustración 57, la idea es destacar las temáticas importantes dependiendo de su incremento en posición, pero sobre todo encontrar una explicación a esto. En la parte de arriba a la izquierda, se ve que para cada fuente se pretende decir cuáles son las temáticas que ganaron más en cobertura. Para ello será necesario encontrar una metodología que permita identificar esas temáticas con una señal binaria, basada en una métrica bien definida de lo que representa la tendencia en cada fuente. Esta señal se tendrá que definir para observar lo que sucedió estos últimos 6 meses.

También se puede observar como la relevancia de cada temática es importante para lo que se enseña a los clientes, que se puede relacionar con el requerimiento R1.

Yearly update: Suggested template for the complete view on all issues  
2020 highlight

Summary	Evolution of material issues over the last year				
<p>Noticeable evolution in the sources' coverage of issues</p> <ul style="list-style-type: none"> <li><b>Benchmark:</b></li> <li><b>Soft and hard law:</b></li> <li><b>News:</b></li> <li><b>Social Media:</b></li> </ul> <p>Significant update on material issues</p> <ul style="list-style-type: none"> <li>A significant raise of <b>issue 10</b> which was of high importance in last analysis</li> <li>The emerging nature of <b>issue 1</b> is confirmed</li> <li><b>Issue 4</b> decreases importance in public opinion</li> </ul>	PILAR / CATEGOR Y OF ISSUES #1	Top 3 issues in rank difference	Δ	Importance for the company	Source driving the result and explanation
		Issue 1	+4	Low	Hard law: ...
		Issue 2	-3	High	Soft law: ...
		Issue 3	+2	Medium	News: ...
	PILAR / CATEGOR Y OF ISSUES #2	Top 3 issues in rank difference	Δ	Importance for the company	Source driving the result and explanation
		Issue 4	-3	High	News: ...
		Issue 5	+2	High	Soft law: ...
		Issue 6	+2	Low	Social media: ...
	PILAR / CATEGOR Y OF ISSUES #3	Top 3 issues in rank difference	Δ	Importance for the company	Source driving the result and explanation
		Issue 7	+4	Low	News: ...
		Issue 8	+3	High	Hard law: ...
		Issue 9	-2	Low	Social media: ...
	PILAR / CATEGOR Y OF ISSUES #4	Top 3 issues in rank difference	Δ	Importance for the company	Source driving the result and explanation
		Issue 10	+5	High	Hard law: ...
		Issue 11	4	Low	Hard law: ...
		Issue 12	+2	Medium	News: ...

Ilustración 57 – Plantilla PowerPoint de la responsable de customer success, para que puedan comunicar los cambios importantes. Diapositiva de evolución de las temáticas. (Fuente: Datamaran)

Esta diapositiva corresponde al requerimiento R11 (para la parte de arriba a la izquierda) y el detalle de cada temática (parte derecha) corresponde al requerimiento R13. En el último, se pretende ser capaz de ir hasta encontrar los artículos, textos de ley, informes corporativos que provocaron el cambio en la temática, para identificar mejor el impacto de ciertos eventos. Varios clientes reportan que están muy interesados por el módulo de *monitoring* pero que siempre se quedan preguntando: ¿Por qué esta temática está ganando en importancia? Y esta pregunta se queda la mayoría de las veces sin respuesta o con una respuesta no satisfactoria por que el usuario espera una respuesta en una frase que resume el evento mayor que sucedió. Por ejemplo: ¿Por qué en abril 2020 mi temática *Climate change & air Quality* ganó 3 posiciones? La respuesta es probablemente porque salieron muchos artículos sobre cómo la pandemia había reducido la contaminación del aire durante el encerramiento.

Para ello es importante construir una metodología que al menos tenga sentido matemático y consiga resultados significativos para el usuario.

La Ilustración 58 presenta un resumen del tipo de *insights* que se puede obtener detectando ciertas señales. Esta tabla tiene como objetivo dar indicaciones sobre el estado actual de una temática, usando un indicador que puede probablemente significar que la temática seguirá ganando en importancia en el futuro. Automatizar la creación de aquella tabla significa cumplir el requerimiento R12.

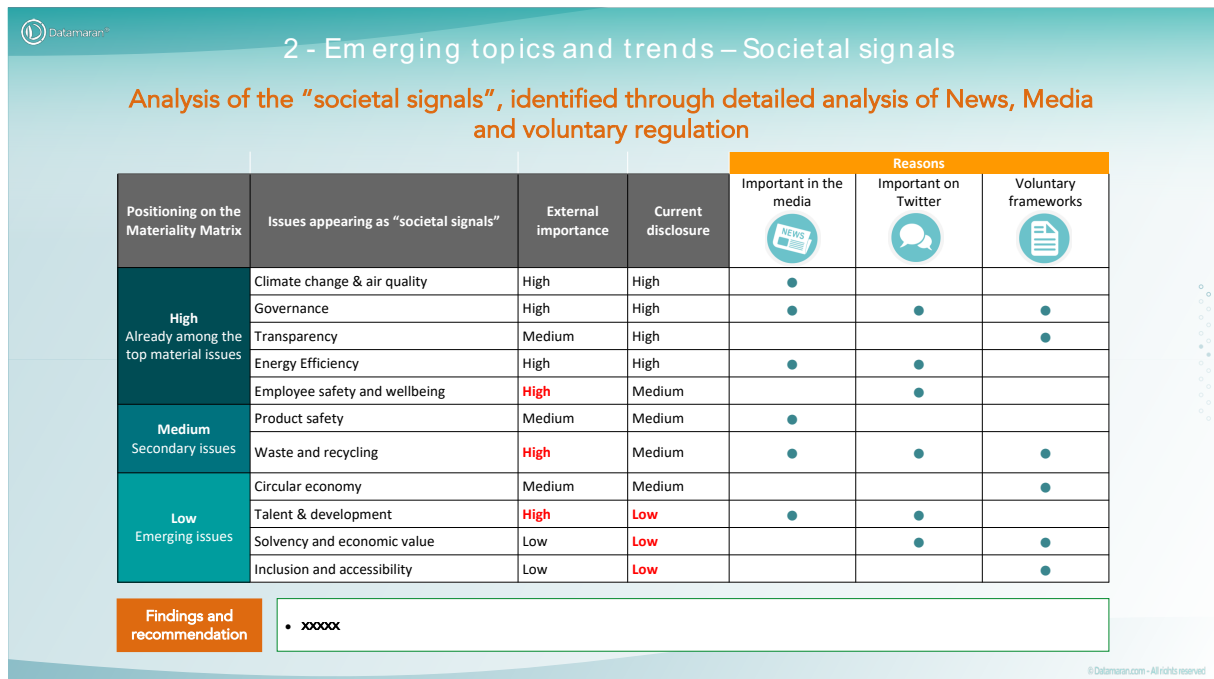


Ilustración 58 - Diapositiva presentando las señales de importancia social atribuida a ciertas temáticas. (Fuente: Datamaran).

## 4.5. Conclusión

En este proyecto se pueden destacar dos clientes principales, el departamento responsable del producto de Datamaran, y el departamento responsable de la experiencia de los usuarios. Estos dos departamentos serán los que enunciarán los requerimientos del nuevo módulo. Ellos representarán la visión del usuario final, ya que son los que tienen contacto regular con ellos y saben formular las necesidades de los clientes. Segundo, el departamento técnico de Datamaran aparece como cliente secundario, porque el remodelado de la plataforma también se motiva por una necesidad de desempeño global del software.

Estos clientes expresaron sus necesidades, la mayoría siendo bien determinadas, sin mucha duda sobre la manera de ejecutarlas (funcionalidades, experiencia visual, arquitectura), pero ciertos requerimientos se dejaron más abiertos y necesitaran una investigación profunda de métodos y de datos para tratarlos correctamente.

En resumen, es necesario definir 3 indicadores:

1. Un indicador de que la temática está de moda
2. Un indicador de que la temática parece que ganará importancia en el futuro
3. Un indicador que permita determinar si un evento dado permite explicar un incremento de importancia de una temática.

Con esos indicadores se podrá pretender automatizar un diagnóstico del análisis del cliente y facilitarle más *insights* posibles para ayudar su toma de decisión y su gestión de riesgos

## 5. Definición de los indicadores

### 5.1. Introducción

Hasta aquí se ha detallado el funcionamiento de Datamaran 4, la versión actual del *software* en línea de la empresa Datamaran. Este trabajo trata de presentar un proyecto de mejora de un módulo específico de la plataforma: el *continuous monitoring*. En el capítulo anterior se han definido los requerimientos del proyecto: algunos no incluyen ninguna libertad interesante a la hora de cumplirlos, pero los 3 últimos R11, R12 y R13 requieren más creatividad. En efecto, estos requerimientos implican la definición de un método de detección de ciertas tendencias, propiedades de un análisis dado en la plataforma.

En este capítulo será necesario estudiar los datos y crear una metodología que permitirá identificar las temáticas que tienen propiedades interesantes para el usuario. Más específicamente los requerimientos R11 y R12 se resolverán introduciendo 2 indicadores (uno para cada requerimiento). El indicador será una señal, que se activará para una temática cuando aquella cumpla una propiedad matemática que tendrá que tratar de reflejar de la mejor manera posible el requerimiento que se desea cumplir. No se presentará el trabajo del requerimiento R13, porque no entra en el enfoque de este trabajo, pero será una futura línea de trabajo.

La estructura del capítulo será la siguiente: primero se explicará la metodología a seguir para definir los indicadores. Luego se detallarán los pasos de aquella metodología para definir el indicador que servirá para cumplir el requerimiento R11. Por fin se realizará un análisis crítico de los resultados obtenidos en ambos indicadores, y se propondrán direcciones de trabajo.

### 5.2. Metodología

Antes de empezar a detallar cómo se van a determinar qué indicadores se usarán para cumplir los requerimientos, y para ello se define una metodología de selección por comparación de alternativas.

#### 5.2.1. Diagrama de flujo

En este apartado se presentará el desarrollo de la metodología empleada, sus distintas etapas, así como una explicación de aquellas.



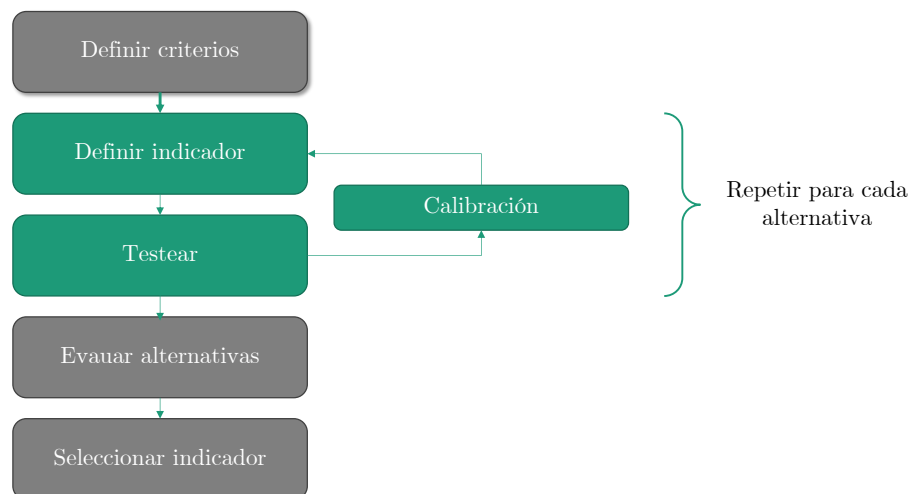


Ilustración 59 - Diagrama de flujo del proceso de selección de un indicador. (Fuente: elaboración propia)

La Ilustración 59 presenta la metodología, cada uno de sus pasos se detallan a continuación.

### Definir criterios

Para cada uno de los dos indicadores que se definirán, será necesario proponer varias soluciones, o alternativas que se comparará entre sí a fin de seleccionar la mejor entre ellas. La “mejor” se determinará según ciertos criterios de selección que se definirán antes de la definición de los criterios. Estos criterios tratarán de tomar en cuenta tanto el desempeño del indicador como el coste de desarrollo de la solución. Los criterios serán diferentes para cada indicador, y se les atribuyen una importancia o peso relativo entre ellos que servirá para comparar los al final.

### Definir indicador

Una vez definidos los criterios de evaluación, empezará el proceso iterativo para crear los indicadores. Para cada indicador se comparará al menos 3 alternativas según criterios medibles y tan objetivos cómo se pueda.

### Testear

Antes de validar un indicador, se enfrenta a una serie de tests (derivados de los criterios definidos) y permitirán tener una mejor idea del desempeño de aquel indicador.

### Calibración

Si el resultado del *test* de desempeño no se aprueba, se hará una iteración sobre los parámetros del indicador, o sobre la definición misma del indicador para intentar incrementar su desempeño hasta que pase los tests.

### Evaluar alternativas

En este momento se dispondrá de 3 alternativas aceptables que podrán servir como indicador, a partir de aquí se tiene que medir el desempeño de cada alternativa para cada uno de los criterios. Con la importancia relativa atribuida a cada criterio, se determinará su peso global de importancia según el método *AHP*<sup>18</sup>. Este método permite construir los pesos globales de una serie de criterios a partir de sus importancias relativas. Dependiendo del peso atribuido a cada uno de los criterios, se combinarán todos los *scores* de la alternativa en cada score con el peso del criterio para construir

<sup>18</sup> Analytic Hierarchy Process: técnica de toma de decisiones en función de la prioridad o jerarquía de criterios que se define.

el score global de la alternativa según la ecuación siguiente:

$$Score(A_i) = \sum_{j \in C} p(C_j) \cdot score(A_i, C_j) \quad (9)$$

$A_i$  es la alternativa número  $i$ .

$C_j$  es el criterio número  $j$  dentro del conjunto  $C$  de todos los criterios

$p(C_j)$  es el peso global del criterio  $C_j$  en %

$score(A_i, C_j)$  es el *score* de la alternativa  $A_i$  para el criterio  $C_j$ .

$Score(A_i)$  es el *score* global de la alternativa  $A_i$

### Seleccionar alternativa

Finalmente, cuando se tiene el score de cada alternativa con todos los criterios tomados en cuenta, se pueden comparar entre sí y la alternativa con el mejor score se seleccionará como solución del problema y respuesta al problema planteado por el requerimiento.

## 5.2.2. Hipótesis

Antes de definir los indicadores que se van a usar, cabe precisar cuáles serán las hipótesis consideradas durante el estudio. Primero, las fuentes (o *stakeholders*) se pueden clasificar según un nivel de reactividad. Se define la reactividad como la capacidad de una fuente de reflejar rápidamente un cambio que ocurre en el mundo real en el análisis de materialidad que se ejecuta. Es decir, cuál es el tiempo medio para cual las consecuencias de un evento del mundo real se pueden observar en un stakeholder. De la más a la menos dinámica:

- Social (o Twitter)
- NewsFlow (noticias de prensa)
- Benchmark
- Recomendaciones legales
- Leyes

Esta clasificación se hace por una parte de manera cualitativa, por lo que se admite de la naturaleza de las fuentes y otra parte más concreta y cuantitativa: por análisis de los datos de las fuentes.

Tabla 12 - Frecuencias de agregación de nuevos elementos a la base de datos. (Fuente: Datamaran)

Fuente	Frecuencia de agregación (elementos/tópico)	Tipo de agregación	Origen cálculo
Social	7/min	Automático/Pipeline	Promedio del volumen total
Newsflow	3/min	Automático/Pipeline	Promedio del volumen total
Benchmark	30/h	Automático/Pipeline	Aproximación
Voluntary Regulations	10/día	Manual	Entrevista con el departamento responsable
Mandatory Regulations	10/día	Manual	Entrevista con el departamento responsable

Se van a estudiar los ritmos de cambio de cada fuente para definir un periodo de tiempo típico de evolución para cada una de ellas.

La Tabla 12 presenta, para cada fuente, una estimación de la frecuencia con la cual se agregan nuevos elementos a la base de datos y se toman en cuenta para los cálculos. El tipo de agregación

permite entender la diferencia entre las fuentes para las cuales la mayoría de los elementos se añaden a mano (textos legales y recomendaciones básicamente) y el resto que es automático. Para lo que es automático, se considera que la base de datos representa en cada instante de manera perfecta el mundo. Esta hipótesis es válida si se considera que el mundo real está bien representado por las diferentes fuentes de datos que Datamaran ha seleccionado. Por ejemplo, no se aceptan todos los artículos de prensa del mundo, obviamente, sino solamente los que se publican en una lista predeterminada de periódicos validados por el equipo. Pero esa selección responde a una necesidad de los usuarios de tener solamente informaciones relevantes para el negocio, por tanto, la base de datos refleja la parte más relevante para los clientes del mundo real.

Para las regulaciones de todos tipos, se ha obtenido una entrevista con la persona encargada del departamento que agrega textos a la base de datos, anotándolos manualmente. La entrevista ha revelado dos cosas interesantes:

- Un cambio en el estilo de gestión del equipo ocurrió a principios de 2019, lo que provocó un incremento significativo del ritmo de agregación,
- Según la responsable, el ritmo actual permite mantener un nivel de representación (en el sentido de adecuación entre el mundo real y la base de datos) “muy decente”.

A partir de aquí, se puede considerar que la representación del mundo real en la base de datos se considerará aceptable. La frecuencia en la Tabla 12 servirá entonces para evaluar cómo evolucionan las fuentes con el tiempo. Cuando el cálculo se ha hecho con el promedio del volumen total significa que se calculó la cantidad promedio de nuevos elementos añadidos sobre un periodo de 6 meses. La comparación entre Social y NewsFlow se hace fácilmente ya que las frecuencias se pueden comparar. Los datos para regulaciones vienen de la entrevista y se presentan tal cual en la Tabla 12. Para *Benchmark* se trata de una estimación: la base de datos se organiza de manera más compleja, ya que los informes se analizan de manera más fina: se cuenta la ratio de mención del tópico en el informe cuando para artículos de prensa solo se indica si el título menciona o no el tópico, por ejemplo. Sabiendo que cada empresa pública sus informes una vez al año, la frecuencia en la base de datos idealmente será de un informe por empresa por año. Se recuerda que se consideran 3 tipos de informes anuales (financiero, de sostenibilidad y emitido por el SEC). Si todas las empresas producen un informe financiero cada año, no todas tienen informe de sostenibilidad o del SEC<sup>19</sup>. Por tanto, en la base de datos cada empresa tiene un promedio de 2 informes por año. La base de datos contiene 5770 empresas distintas, cada informe menciona unos 20 tópicos diferentes en promedio, lo que permite calcular la frecuencia de aproximadamente 30 elementos/tópico/h.

Si estas frecuencias pueden ayudar a entender lo que pasa en volumen en la base de datos, no bastan para entender la dinámica de las fuentes. En efecto, hay muchas cosas que considerar para entender a qué velocidad cambian las posiciones de las temáticas en una fuente, y cada fuente tiene sus propias características que no se detallaron en este trabajo. Por lo tanto, estas estimaciones serán suficientes para justificar la hipótesis sobre la reactividad de las fuentes.

Si las recomendaciones legales se consideran más reactivas que las leyes, es por sentido común: aunque la velocidad de agregación sea la misma, las últimas recomendaciones legales estarán probablemente tratando problemas más actuales que las leyes. Eso lleva a la siguiente hipótesis: los cambios que ocurren en recomendaciones legales se reflejarán tarde o temprano en los textos legales. Los textos de recomendaciones se publican por la mayoría por ONGs, y contienen estudios sobre el estado actual del mundo y tratan de proponer mejores maneras de hacer las cosas. Estos análisis críticos, cuando reciben un apoyo fuerte del público, supuestamente influyen a las entidades legislativas para escribir textos de ley que tratan de comprobar que esas “buenas prácticas” se apliquen. No se ha podido comprobar la validez de la hipótesis con números ya que

---

<sup>19</sup> La Securities and Exchange Commission en los Estados Unidos es la entidad no gubernamental que se encarga de verificar el cumplimiento de las leyes de los mercados financieros.

es un estudio complejo de llevar, pero la hipótesis tiene sentido conceptualmente por lo menos.

Finalmente, la última hipótesis que se hará es sobre la ventana temporal que se observará. Por necesidad del cliente de poder vigilar sus temáticas regularmente, y para ver los cambios más recientes se proporciona una vigilancia sobre 6 meses. De hecho, el registro de conexión de los clientes muestra que, en promedio, cada usuario se conecta una vez al año (generalmente en el momento de redactar su informe corporativo) y otra vez 6 meses después para monitorear los cambios. Se admitirá que esta ventana temporal esté adaptada también para la definición de las señales y su estudio. De manera general se admite que la importancia de una temática se transmite de la fuente menos reactiva a la más dinámica.

## 5.3. Señales sociales

### 5.3.1. Criterios de selección

En este apartado cabe recordar que se quiere cumplir el requerimiento R11: identificar las temáticas de moda. Hoy en día en mi sector. Que para ello se pondrá a disposición del usuario un archivo Excel que se podrá descargar y proveerá una tabla detallando para cada temática sus características. Se quiere definir señales que serán susceptibles de traducir un posible incremento futuro de la importancia de la temática, analizando solamente su estado actual. Se referirá a las temáticas que presentan un indicador por encima del límite de activación de la señal como las temáticas activadas o anunciadas.

Tabla 13 - Criterios de selección de la señal para detectar temáticas de moda actualmente. (Fuente: elaboración propia)

Código criterio	Criterio	Descripción	Valor clave
R11-C1	Fiabilidad	La importancia de las temáticas no anunciadas no incrementa en el futuro	Ratio de fiabilidad, $p(\bar{I} \bar{A})$
R11-C3	Selectividad	Solamente algunas de las temáticas tienen la señal	% de temáticas sin activar, $p(\bar{A})$
R11-C2	Poder predictivo	La importancia de las temáticas anunciadas incrementa en el futuro	Ratio de predicción, $p(I A)$
R11-C4	Económico	El desarrollo de la solución debe ser lo más económico posible	Complejidad de implementación, $\frac{coste}{coste_{min}}$

Entonces, un buen indicador de un futuro incremento de importancia de la temática tiene que verificar los criterios de la Tabla 13. A continuación, se van a detallar la definición de cada criterio presentado en esta tabla. Se emplea una notación de probabilidades condicionales:

$$A, B \in \Omega^2, p(B) > 0: p(A|B) \stackrel{\text{def}}{=} \frac{p(A \cap B)}{p(B)} \quad (10)$$

$p(A)$  y  $p(B)$  son las probabilidades asociadas respectivamente a los eventos  $A$  y  $B$  del universo  $\Omega$ .

$p(A \cap B)$  es la probabilidad asociada a la realización simultánea de  $A$  y  $B$

$p(A|B)$  será la probabilidad del evento  $A$  conocido  $B$

Se puede observar que si  $B$  tiene una probabilidad nula,  $p(A|B)$  no está definido, con lo cual por defecto se tomará  $p(A|B) = 0$  si  $p(B) = 0$ .

### Fiabilidad

La identificación de temáticas que incrementarán potencialmente tiene que ser precisa: las temáticas que se descartan tienen que ser las que no son “interesantes”, significando que estas no van a ganar posiciones en el futuro próximo. Por lo menos, la ratio de temáticas correctamente descartadas debe de ser máximo, y se define con la siguiente ecuación:

$$R_{fiabilidad}(\%) = \frac{N_{no\ anunciado\ y\ no\ incrementando}}{N_{total\ temáticas} - N_{anunciado}} \approx p(\bar{I}|\bar{A}) \quad (11)$$

$R_{fiabilidad}$  es la ratio de fiabilidad del indicador

$N_{anunciado}$  es el número de temáticas que se activan en un mes  $m$

$N_{no\ anunciado\ y\ no\ incrementando}$  es el número de temáticas dentro de las no anunciadas que realmente no ganaron en importancia entre los meses  $m$  y  $m + 6$

$I$  es el evento: “la temática incrementa su ranking en los próximos 6 meses”

$A$  es el evento: “la temática está activada”

### Selectividad

Otro criterio importante desde el punto del usuario es la selectividad: aunque se quiere un indicador que funcione, tampoco se quiere un indicador que anuncie la mitad de las temáticas como incrementando porque pierde valor para el usuario. La selectividad se mide:

$$R_{selectividad}(\%) = 1 - \frac{N_{anunciado}}{N_{total\ temáticas}} \approx p(\bar{A}) \quad (12)$$

$R_{selectividad}$  es la ratio de selectividad del indicador.

$N_{anunciado}$  es el número total de temáticas activadas en un mes  $m$

$N_{total\ temáticas}$  es el número total de temáticas definidas por el análisis del usuario

$A$  es el evento: “la temática está activada”

### Poder predictivo

No solo se quiere un indicador fiable para su predicción negativa (el hecho de descartar las temáticas no interesantes), pero que también sea capaz de anunciar las que incrementarán. Es decir, si hoy una temática hubiera ganado en importancia, un buen indicador lo habría enunciado 6 meses antes. Se quiere el máximo ratio de sensibilidad:

$$R_{predicción}(\%) = \frac{N_{confirmado}}{N_{anunciados}} \approx p(I|A) \quad (13)$$

$R_{predicción}$  es la ratio de predicción del indicador.

$N_{anunciados}$  es el número total de temáticas que se activan con la señal para el análisis

$N_{confirmado}$  es el número de temáticas dentro de las anunciadas que realmente ganaron en importancia entre los meses  $m$  y  $m + 6$

$I$  es el evento: “la temática incrementa su ranking en los próximos 6 meses”

$A$  es el evento: “la temática está activada”

### Económico

Los recursos necesarios para el desarrollo de la solución también importan a la hora de seleccionar un indicador. Este criterio será evaluado de manera subjetiva con una estimación del tiempo de desarrollo que se necesitaría.

## Coherencia

Este criterio toma en cuenta la facilidad con la cual el usuario puede aceptar o no la solución, si parece sensato definir una señal así o no. También será subjetivo en función de cómo se define la señal y de la justificación que viene detrás.

Para cada una de las alternativas definidas, se evaluarán estos indicadores con una puntuación entre 0 y 100. El *score* global de cada alternativa será una media ponderada de estas puntuaciones, el peso de cada uno de los criterios se van a definir a continuación usando el método AHP<sup>20</sup>. Este método se basa en la jerarquía (o importancia relativa de los criterios 2 a 2) para construir un vector de pesos globales para cada criterio. Una vez estos pesos calculados, se puede evaluar un *score* global para cada alternativa, como por ejemplo en la Tabla 14.

Tabla 14 - Ejemplo de evaluación de una alternativa. (Fuente: elaboración propia)

Criterios	Alternativa 1		
	Puntuación	Peso	Score ponderado
Fiabilidad	71%	49%	10%
Selectividad	96%	19%	3%
Poder predictivo	20%	19%	5%
Coherencia	50%	9%	5%
Económico	100%	4%	16%
<b>Total</b>		<b>100%</b>	<b>65.16%</b>

Para determinar aquellos pesos, necesitamos como mínimo 3 comparaciones 2 a 2 de criterios, usando una escala cuantitativa de evaluación subjetiva, como aquella de la Tabla 15.

Tabla 15 - Escala de medida de importancia relativa de criterios. (Fuente: (Saaty, 2008))

Escala	Definición	Explicación
1	Igualmente, preferida	Los dos criterios contribuyen igual al objetivo
3	Moderadamente preferida	La experiencia y el juicio favorecen un poco a un criterio frente al otro
5	Fuertemente preferida	La experiencia y el juicio favorecen fuertemente a un criterio frente al otro
7	Muy fuertemente preferida	un criterio es favorecido muy fuertemente sobre el otro. En la práctica se puede demostrar su dominio
9	Extremadamente preferida	La evidencia favorece en la más alta medida a un factor frente al otro

Para construir la matriz correspondiente, y entonces deducir el vector de pesos que mejor representará el objetivo que se quiere cumplir, se necesitan 3 comparaciones de criterios independientes:

- Fiabilidad es moderadamente preferida a selectividad, porque es el criterio más importante para el usuario: que lo que se descarta sea realmente poco importante
- Selectividad es igual de preferida que poder predictivo.

<sup>20</sup> AHP *Analytic Hierarchy Process* véase (Saaty, 2008)

- Poder predictivo es entre igual y moderadamente preferida a coherencia
- Coherencia es moderadamente preferida a económico, porque los costes de implantación en complejidad suelen ser realmente despreciables comparado con la justificación de la metodología.
- Fiabilidad es extremadamente preferida a económico, porque los costes de implantación en complejidad suelen ser realmente despreciables comparado con la justificación de la metodología.

La matriz de que se construye referenciando en cada celda la importancia relativa del criterio del encabezado de línea sobre el criterio de encabezado de columna. Según el método AHP se obtiene el vector de prioridad visible en la Tabla 16.

Tabla 16 - Matriz de prioridad de los criterios. En negrita están los números entrados manualmente, el resto se calcula automáticamente de forma coherente. (Fuente: elaboración propia)

	Fiabilidad	Selectividad	Poder predictivo	Coherencia	Económico	Prioridad
Fiabilidad	1.00	<b>3.00</b>	3.00	6.00	<b>9.00</b>	49%
Selectividad	0.33	1.00	<b>1.00</b>	2.00	6.00	19%
Poder predictivo	0.33	1.00	1.00	<b>2.00</b>	6.00	19%
Coherencia	0.17	0.50	0.50	1.00	<b>3.00</b>	9%
Económico	0.11	0.17	0.17	0.33	1.00	4%

Tabla 17 - Estimaciones del valor propio máximo de la matriz de jerarquía. (Fuente: elaboración propia)

	$\lambda_{max}$
Fiabilidad	5.11
Selectividad	5.02
Poder predictivo	5.02
Coherencia	5.02
Económico	5.13

En la Tabla 17 se presentan las estimaciones del valor propio máximo  $\lambda_{max}$  de la matriz de la Tabla 16. El método AHP sugiere coger el promedio de las estimaciones y calcular la ratio de consistencia de acuerdo con la fórmula:

$$Ratio\ consistencia = \frac{\widehat{\lambda_{max}} - n}{n - 1} \cdot \frac{1}{RI} \quad (14)$$

$\widehat{\lambda_{max}}$  es el estimador (promedio aquí) de  $\lambda_{max}$

$n$  es la dimensión de la matriz de jerarquía ( $n = 5$  en este caso)

$RI$  es la ratio de consistencia de una matriz aleatoria de dimensión  $n$  dado por la Tabla 18

Tabla 18 -Ratio de consistencia de una matriz aleatoria en función de su dimensión. (Fuente: (Saaty, 2008))

Dimensión n	1	2	3	4	5	6
RI	0	0	0.58	0.9	<b>1.12</b>	1.24

Según el trabajo de Thomas L. Saaty (Saaty, 2008), la ratio de consistencia debe de ser  $< 10\%$  para que la jerarquía sea aceptable, el cálculo de la ecuación (14) da  $RI = 2\%$ , con lo cual se



acepta esta definición de los pesos de los criterios según el método AHP.

### 5.3.2. Definición de las alternativas

#### Alternativa 1 – Opinión pública contra el resto

Una primera idea cuando se pretende medir, dada la situación actual, la probabilidad para una temática de ganar en importancia en el futuro, es basarse en las fuentes más dinámicas y compararlas a las otras fuentes. El indicador para una temática estará activado si la temática está considerada más importante al menos en una de las fuentes muy reactivas (redes sociales, artículos de prensa) que en todas las fuentes lentas. La condición matemática para la activación será:

$$\max_{f \in \text{fuentes dinámicas}} \text{score}(T_i, f) > \max_{f \in \text{fuentes lentas}} \text{score}(T_i, f) + t_{10} \quad (15)$$

$T_i$  es la temática número  $i$  dentro del conjunto de temáticas.

$\text{score}(T_i, f)$  es el score de  $T_i$  en la fuente  $f$  dentro del conjunto de todas las fuentes de Datamaran (*benchmark, newsflow, social, mandatory regulations, voluntary regulations*).

*fuentes dinámicas* se refiere al conjunto de las fuentes más dinámicas según lo que se detalla en el apartado 5.2.2 Hipótesis. Serán las fuentes *newsflow*, y *social*.

*fuentes lentas* se refiere al resto de fuentes

$t$  será un parámetro de la alternativa, será el valor límite de la diferencia para activar la señal

Se recuerda que el score se refiere al *percentile rank* – una transformación de la clasificación que atribuye un valor normalizado (entre 0 y 1) a cada temática. El parámetro  $t$  se determinará en el apartado de selección de la alternativa, optimizando su valor para obtener el mejor score para esta alternativa.

#### Alternativa 2 – Indicador granular con varios subindicadores

Este indicador se basa sobre la misma observación que el primero: la diferencia de dinamismo entre las fuentes. El objetivo es crear un indicador que refleje lo que se puede esperar del comportamiento de las fuentes: una transmisión de la información desde lo más reactivo hacia los menos reactivos. Para ello se definirán 4 subindicadores, relacionados con 4 grados de libertad del análisis:

- Newsflow: comparar la importancia de la temática en artículos de prensa con su importancia en *Benchmark*. Se pretende que una temática que es más importante hoy en día para la prensa tendrá que impactar lo que se redacta en los informes corporativos. Como pueden existir diferentes enfoques en *benchmark*, se calculará una media ponderada de este resultado. El subindicador se activa si:

$$\text{score}(T_i, \text{news}) > \max_{b \in \text{benchmarks}} \text{score}(T_i, b) + t_{20} \quad (16)$$

- Social: lo mismo que con *NewsFlow*, pero se compara Twitter a todas las otras fuentes, para que el criterio sea más selectivo, ya que Twitter es un media en general menos fiable que la prensa. Como pueden existir diferentes enfoques en *benchmark*, se calculará una media ponderada de este resultado. El subindicador se activa si:

$$\text{score}(T_i, \text{social}) > \max_{f \in \text{fuentes lentas}} \text{score}(T_i, f) + t_{21} \quad (17)$$

- Leyes: se compara la importancia en recomendaciones legales a la importancia en leyes de la temática. Se basa en la hipótesis de transferencia de las medidas desde las recomendaciones a la aplicación firme de regulaciones con textos de ley. El subindicador se activa si:

$$\text{score}(T_i, \text{recomendaciones}) > \text{score}(T_i, \text{leyes}) + t_{22} \quad (18)$$

- Alineación interna/externa: Datamaran deja a sus usuarios entrar manualmente como ellos alojan recursos para gestionar las temáticas como ellos piensan que se tiene que hacer en su entorno. Uno de los objetivos de este ejercicio es llevarlos a darse cuenta de las temáticas que son más importantes de lo que ellos creen, y tratar de poner el foco en estas.

$$score(T_i, interno) > stakeholder score + t_{23} \quad (19)$$

### Alternativa 3 – Inferencia de las reglas a partir del conjunto de datos colectados

Para testear la validez de los indicadores, se usará una base de datos conteniendo 15 empresas realizando varios análisis sobre su entorno. En total, siguiendo estos análisis varios meses para observar su evolución se consigue un conjunto de datos conteniendo 75 puntos de comparación sobre una evolución real de 6 meses en los indicadores. Este conjunto contiene datos reales de empresas representativas de los clientes de Datamaran y análisis definidos correctamente. Una idea también es observar esta base de datos y sacar conclusiones sobre los datos característicos de las temáticas que ganan importancia y por inferencia, definir señales características que permiten identificar las temáticas sensibles con antelación.

Se pueden usar varias combinaciones de fuentes para llegar a definir una señal a partir de los datos, aquí se comparará *benchmark* y *stakeholder score* (el *score* global de la temática). La condición de la señal activada será:

$$score(T_i, benchmark) > stakeholder score(T_i) + t \quad (20)$$

### 5.3.3. Evaluación y selección

Una vez los criterios y las alternativas bien definidos, se tienen que evaluar una primera vez y luego iterar sobre los parámetros (las  $t$  en las definiciones) para encontrar una definición que mejor cuadra con el objetivo deseado. Se optimizan los parámetros de cada una de las alternativas con una exploración de las posibilidades de forma determinística e individual de los parámetros. Para las alternativas de un solo parámetro se itera paso a paso entre 0 y 1 y se guarda la mejor solución entre todas. Para la alternativa 2, se usó el mismo método para cada parámetro considerado independientemente de los demás. Se obtienen los resultados siguientes presentados en la Tabla 19.

Tabla 19 - Resultados de la optimización de los parámetros con una búsqueda local. (Fuente: elaboración propia)

$t$	Alternativa 1	Alternativa 2	Alternativa 3
$t_{i0}$	0.12	0.1	0.12
$t_{i1}$		0.03	
$t_{i2}$		0.03	
$t_{i3}$		0	

Tabla 20 - Resultados obtenidos por las 3 alternativas según los criterios y su respectivo peso. (Fuente: elaboración propia)

Criterios	Peso	Alternativa 1		Alternativa 2		Alternativa 3	
		Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	71%	35%	70%	35%	75%	37%
Selectividad	19%	97%	18%	90%	17%	72%	14%
Poder predictivo	19%	27%	5%	28%	5%	37%	7%
Coherencia	9%	60%	6%	100%	9%	20%	2%
Económico	4%	100%	4%	90%	3%	100%	4%
<b>Total</b>	<b>100%</b>		<b>68%</b>		<b>70%</b>		<b>63%</b>

Los resultados de la Tabla 20 se obtienen (para los criterios: fiabilidad, selectividad y poder predictivo) a partir del conjunto de datos del estudio. La coherencia se estima así porque:

- La alternativa 1 obtiene un 60% de coherencia porque el indicador es demasiado sencillo comparado con la complejidad de la plataforma, no encaja bien con los métodos generales de Datamaran. No obstante, comparar las fuentes sociales con el resto tiene sentido con lo que se espera de estas señales.
- La alternativa 2 obtiene el máximo de puntos, porque es la más coherente, una granularidad en las señales que estudian fuentes interesantes y coherentes con la sociedad en un momento dado.
- La alternativa 3 es la menos coherente, compara *benchmark* con el *stakeholder score* es más difícilmente justificable, puede que sea una realidad en los datos, no tiene una base sólida a la hora de explicarlo.

Por el criterio económico:

- La alternativa 1 es la más sencilla de implantar, el indicador es único y necesita pocos datos para generarse
- La alternativa 2 también es sencilla, pero se tienen que efectuar los cálculos con 4 subindicadores y calcular la suma de los activados, lo que será ligeramente más complejo en el código.
- La alternativa 3 recibe también un 100% de puntuación como la primera porque solo será un cálculo sencillo.

Dados los resultados generales, se seleccionará la opción definida con la alternativa 2, la solución se implementará en una hoja de cálculo Excel generada automáticamente a partir del código en Python, y la hoja principal se presentará como en la Ilustración 60.

**[SOCIETAL SIGNALS TO CONSIDER] - EXPLANATION**

This analytical lens should complement your annual or bi-annual materiality review.

It is based on the data of your analysis as of 04-2020, and provides an additional interpretation of the results by ignoring the source weights applied.

Each column represents a comparison of one or more leading vs. lagging source scores as indicated in the column headers. If the leading source has a higher score, the cell value is TRUE.

Issues with a 'Total' value of 3 or 4 present stronger signals for consideration; 2 is weaker; 1 and 0 are not presenting.

SOCIETAL SIGNALS TO CONSIDER					
Issue name	External score > Internal score	Newsflow > Benchmark(s)	Voluntary regulation > Mandatory regulation	Social > All other sources	Total
Biodiversity	TRUE	FALSE	FALSE	FALSE	1
Business Ethics	TRUE	FALSE	TRUE	FALSE	2
Climate & Energy	FALSE	TRUE	TRUE	FALSE	2
Digitalization & Data Security	FALSE	FALSE	FALSE	FALSE	0
Disclosure	TRUE	FALSE	TRUE	FALSE	2
Customer Engagemant	FALSE	FALSE	FALSE	TRUE	1
Emissions (Air & Soil)	TRUE	FALSE	FALSE	FALSE	1
Employee Engagement & Diversity	FALSE	FALSE	FALSE	FALSE	0
Geopolitical Risk	TRUE	TRUE	FALSE	FALSE	2
Governance Structure	TRUE	FALSE	TRUE	TRUE	3
Health & Safety/Product Stewardship	TRUE	FALSE	FALSE	FALSE	1
Human Rights	TRUE	TRUE	FALSE	TRUE	3
Innovation	FALSE	FALSE	FALSE	FALSE	0
Organizational Resilience	TRUE	FALSE	TRUE	FALSE	2
Responsible Supply Chain	FALSE	FALSE	FALSE	TRUE	1
Stakeholder Engagement	FALSE	TRUE	TRUE	FALSE	2
Sustainable Finance	FALSE	TRUE	TRUE	FALSE	2
Resource Efficiency & Waste	TRUE	TRUE	FALSE	FALSE	2
Water	TRUE	FALSE	FALSE	FALSE	1

© Datamaran. All rights reserved - www.datamaran.com

Summary Emerging Issues Analysis +

Ilustración 60 - Ejemplo de diagnóstico estático detectando las temáticas que probablemente ganarán en importancia en los próximos 6 meses según el método elegido. (Fuente: elaboración propia)

### 5.3.4. Crítica de la selección

Primero, se puede observar que los resultados obtenidos dependen fuertemente de los criterios subjetivos (económico y coherencia) aunque no tengan un peso mayor. Si se consideran sólo los criterios cuantitativos se podría seleccionar una alternativa diferente, pero la coherencia de la solución varía significativamente entre ellas. En términos de desempeño puro (mirando solo a los 3 primeros criterios), no hay una alternativa que domine las demás en el sentido de Pareto<sup>21</sup>. Se puede cuestionar la utilidad de estas definiciones entonces, comparándolas con lo que pasaría en dos situaciones triviales.

Tabla 21 - Puntuación obtenida por dos alternativas triviales. (Fuente: elaboración propia)

Criterios	Peso	Anunciar todas		No anunciar ninguna	
		Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	0%	0%	71%	35%
Selectividad	19%	0%	0%	100%	19%
Poder predictivo	19%	29%	5%	0%	0%
Coherencia	9%	0%	0%	0%	0%
Económico	4%	100%	4%	100%	4%
<b>Total</b>	<b>100%</b>		<b>9%</b>		<b>57%</b>

La Tabla 21 demuestra que anunciar todas las temáticas no daría buenos resultados (la fiabilidad será nula ya que no se descarta ninguna). En ambos casos la coherencia será 0 (imposible convencer a un cliente de que eso es una buena metodología de detección de las temáticas importantes), y el criterio económico siempre será 100 (porque no cuesta nada implementar la

<sup>21</sup> Una alternativa domina si presenta una mejor puntuación que las demás en todos los criterios

solución). No obstante, la opción de no anunciar nada obtiene resultados decentes, porque la fiabilidad natural (la probabilidad que una temática no gane posición en el ranking) es bastante alta (71%). Los resultados muestran que los indicadores son capaces de mantener el nivel de fiabilidad natural (o mejorarlo muy poco) incluso seleccionando algunas temáticas y eso será el valor añadido para el usuario. La conclusión es que la mejor solución para el usuario será la solución que está esperando, y que con la cual se sentirá más cómodo, así tendrá más confianza en los resultados.

Es importante destacar que, si el método para encontrar la alternativa 3 se basa en un análisis de los datos que sirven para validar las alternativas, puede ocurrir que los pesos aplicados no sean compatibles con lo que se descubre. Fue el caso en este estudio y demuestra que es importante.

## 5.4. Tendencia

Además de un diagnóstico del estado de una temática en un instante dado, se requiere analizar las informaciones sobre su evolución en los últimos 6 meses, para detectar tendencias interesantes. Como para el apartado anterior, se quiere sobre todo descartar las temáticas que no son interesantes, es decir que probablemente no ganarán en importancia en los próximos meses. Para ello se requiere analizar las fuentes que se consideran como motrices de cambio: *Twitter*, *Newsflow* y recomendaciones legales. Para cada temática se analizará su importancia de aquella dentro del conjunto de temáticas, en vez de usar el *score* (o *percentile rank*) relacionado con la posición de la temática en el ranking, se usará lo que se llama la **presencia** de la temática. La presencia se define:

$$P(T_i, f) \stackrel{\text{def}}{=} \frac{V(T_i, f)}{\sum_{k \in I} V(T_k, f)} \quad (21)$$

$P(T_i, f)$  y  $V(T_i, f)$  son la presencia y el volumen (respectivamente) de la temática  $T_i$  en la fuente  $f$

$i \in I$  el conjunto de todos los índices de temáticas del análisis.

El uso de la presencia como medida permite librarse de las restricciones del ranking que son:

- No hay nada encima del rango 1, una temática en posición 1 ya no puede ganar puntos
- Se pierde la información de distancias relativas entre temáticas

También se podrían evitar estas restricciones usando un volumen bruto de datos (ej. el número absoluto de artículos de prensa mencionando una temática), pero este volumen depende mucho de la gestión interna de la base de datos de Datamaran. Si un día Datamaran decide incluir más fuentes de artículos de prensa (es decir vigilar más periódicos para leer todos sus artículos), el volumen en todas las temáticas aumentará probablemente conforme aumenta el volumen total, pero eso no aporta información sobre las temáticas en sí. Lo que sirve al usuario es saber si una temática, dado su evolución temporal últimamente, parece seguir una forma de tendencia que se confirmará en los meses siguientes. Para este problema también cabe decir que la predicción (sobre todo para fenómenos sociales complejos como es el caso aquí) es una disciplina muy complicada y que para que las mejores tecnologías de predicción funcionen se necesita una gran cantidad de datos (más que lo que se tiene para este trabajo). El objetivo será clasificar las temáticas según una lógica definida para intentar estimar los futuros incrementos, para ello se observarán más específicamente las 3 fuentes dinámicas que son: *Twitter*, artículos de prensa y recomendaciones legales. Se analizarán las tendencias en esas 3 fuentes por separado y se calculará el número de fuentes presentando una señal para ayudar al usuario a rápidamente ubicar las temáticas más relevantes de las demás.

Para la realización de este estudio de desempeño y selección de alternativas, se usará un conjunto de datos reuniendo 17 empresas y un total de 416 temáticas diferentes en su evolución durante 10 meses. Este número está bloqueado por parte porque los cambios en la definición de

los tópicos en 2019 impiden analizar correctamente los datos anteriores a esta fecha, pero más tarde en la empresa se extenderá este estudio a un conjunto de datos más amplio.

### 5.4.1. Criterios de selección

Para cada una de las fuentes que se van a evaluar se seleccionará una alternativa entre las que se van a presentar para la definición de los indicadores. Los criterios de selección serán esencialmente los mismos que para el apartado 5.3.1, pero definidos sobre cantidades diferentes. Los datos usados para analizar una temática se representan en la línea ejemplo:

Tabla 22 - Ejemplo de datos usados para evaluar el desempeño de las señales. Para cada mes se usa la presencia de la temática (%). (Fuente: Datamaran)

La Tabla 22 presenta los datos típicos de una línea de la tabla completa, para cada mes se usa la presencia (%) definida en la ecuación (21), luego la columna “Incrementa 3 meses después” valdrá *TRUE* si la presencia aumenta entre los meses 6 y 9, *FALSE* si no. Finalmente, la columna “Análisis” contiene el nombre de la empresa que realizó el análisis de esta temática.

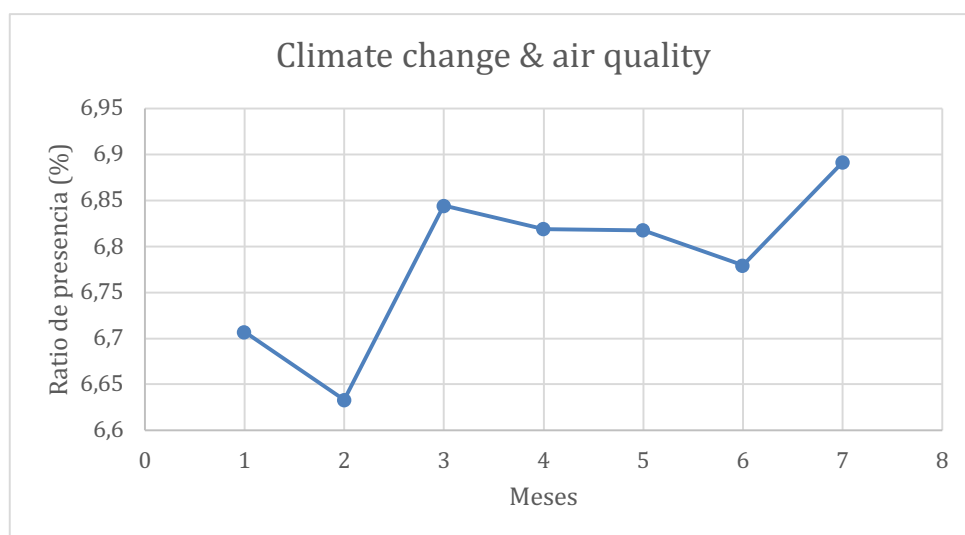


Ilustración 61 - Gráfico representando la evolución temporal de la presencia de la temática Climate change & air quality. (Fuente: elaboración propia)

La Ilustración 61 muestra un ejemplo de evolución de los datos con el tiempo, no se ponen fechas porque el análisis debe de hacerse relativamente a la fecha de inicio del estudio, no importa la fecha. Además, se sabe que esta temática seguirá incrementando globalmente su presencia en los 3 meses siguientes.

Temática	Mes 0	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Incremento 3 meses después	Análisis
Climate change & air quality	11.5	11.5	11.5	12.5	13.4	13.0	12.7	TRUE	Amazon

#### Fiabilidad

Con la activación de señales se quiere sobre todo que haya lo menos posible falsos negativos: temáticas que no se anuncian pero que realmente ganan en importancia en los meses siguientes:

$$R_{fiabilidad}(\%) = \frac{N_{no\ anunciado\ y\ no\ incrementando}}{N_{total\ temáticas} - N_{anunciado}} \approx p(\bar{I}|\bar{A}) \quad (22)$$

$R_{fiabilidad}$  es la ratio de fiabilidad del indicador

$N_{anunciado}$  es el número de temáticas que se activan

$N_{no\ anunciado\ y\ no\ incrementando}$  es el número de temáticas dentro de las no anunciadas que realmente no ganaron en importancia entre los meses 6 y 9

$I$  es el evento: “la temática incrementa su presencia entre los meses 6 y 9”

$A$  es el evento: “la temática está activada”

### Selectividad

Otro criterio importante desde el punto del usuario es la selectividad: aunque se quiere un indicador que funcione, tampoco se quiere un indicador que anuncie la mitad de las temáticas como incrementando porque pierde valor para el usuario. La selectividad se mide:

$$R_{selectividad}(\%) = 1 - \frac{N_{anunciado}}{N_{total\ temáticas}} \approx p(\bar{A}) \quad (23)$$

$R_{selectividad}$  es la ratio de selectividad del indicador.

$N_{anunciado}$  es el número total de temáticas activadas en un mes dado

$N_{total\ temáticas}$  es el número total de temáticas definidas por el análisis del usuario

$A$  es el evento: “la temática está activada”

### Poder predictivo

No solo se quiere un indicador fiable para su predicción negativa (el hecho de descartar las temáticas no interesantes), pero que también sea capaz de anunciar las que incrementarán. Es decir, si hoy una temática hubiera ganado en presencia, un buen indicador lo habría enunciado 3 meses antes. Se quiere el máximo ratio de sensibilidad:

$$R_{predicción}(\%) = \frac{N_{confirmado}}{N_{anunciados}} \approx p(I|A) \quad (24)$$

$R_{predicción}$  es la ratio de predicción del indicador.

$N_{anunciados}$  es el número total de temáticas que se activan con la señal para el análisis

$N_{confirmado}$  es el número de temáticas dentro de las anunciadas que realmente ganaron en importancia entre los meses 6 y 9

$I$  es el evento: “la temática incrementa su presencia en los próximos 3 meses”

$A$  es el evento: “la temática está activada”

### Económico

Los recursos necesarios para el desarrollo de la solución también importan a la hora de seleccionar un indicador. Este criterio será evaluado de manera subjetiva con una estimación del tiempo de desarrollo que se necesitaría.

### Coherencia

Este criterio toma en cuenta la facilidad con la cual el usuario puede aceptar o no la solución, sí parece sensato definir una señal así o no. También será subjetivo en función de cómo se define la señal y de la justificación que viene detrás.



Para cada una de las alternativas definidas, se evaluarán estos indicadores con una puntuación entre 0 y 100. El *score* global de cada alternativa será una media ponderada de estas puntuaciones, el peso de cada uno de los criterios se va a definir a continuación usando el método AHP.

Para construir la matriz correspondiente, y entonces deducir el vector de pesos que mejor representará el objetivo que se quiere cumplir, se necesitan 3 comparaciones de criterios independientes:

- Fiabilidad es moderadamente preferida a selectividad, porque es el criterio más importante para el usuario: que lo que se descarta sea realmente poco importante
- Selectividad es igual de preferida que poder predictivo.
- Poder predictivo es entre igual y moderadamente preferida a coherencia
- Coherencia es moderadamente preferida a económico, porque los costes de implantación en complejidad suelen ser realmente despreciables comparado con la justificación de la metodología.
- Fiabilidad es extremadamente preferida a económico, porque los costes de implantación en complejidad suelen ser realmente despreciables comparado con la justificación de la metodología.

La matriz de que se construye referenciando en cada celda la importancia relativa del criterio del encabezado de línea sobre el criterio de encabezado de columna. Según el método AHP se obtiene el vector de prioridad visible en la Tabla 16.

Tabla 23 - Matriz de prioridad de los criterios. En negrita están los números entrados manualmente, el resto se calcula automáticamente de forma coherente. (Fuente: elaboración propia)

	Fiabilidad	Selectividad	Poder predictivo	Coherencia	Económico	Prioridad
Fiabilidad	1.00	<b>3.00</b>	3.00	6.00	<b>9.00</b>	49%
Selectividad	0.33	1.00	<b>1.00</b>	2.00	6.00	19%
Poder predictivo	0.33	1.00	1.00	<b>2.00</b>	6.00	19%
Coherencia	0.17	0.50	0.50	1.00	<b>3.00</b>	9%
Económico	0.11	0.17	0.17	0.33	1.00	4%

De la misma manera que en el apartado 5.3.1 se comprueba la buena coherencia de la matriz de jerarquía y se valida el vector de pesos definido en la Tabla 16.

## 5.4.2. Definición de las alternativas

### Alternativa 1 – Diferencia sencilla

Una primera idea sencilla para intentar predecir una tendencia es calcular el incremento global sobre 6 meses:

- si este incremento es positivo, es posible que sea una tendencia que va a continuar en los meses siguientes y entonces provocar un incremento 3 meses después
- si este incremento es negativo, se puede admitir que la temática no está en periodo de tendencia, y no incrementará en presencia en el futuro próximo.

La condición de activación de la temática en este caso será:

$$P(T_i, m) - P(T_i, m - 6) > t_1 \quad (25)$$

$P(T_i, m)$  es la presencia de la temática  $T_i$  durante el mes  $m$

$t_1$  es el parámetro de corrección que permitirá ajustar la definición de la señal para optimizar sus resultados

### Alternativa 2 – Desviación estándar

Observando los datos, se puede ver que algunas temáticas varían más que otras, con lo cual la solución anterior (que fija un límite que la diferencia tiene que superar) no toma en cuenta las especificidades de cada temática. La idea de esta alternativa es identificar las características propias de las temáticas, por ejemplo, usando la presencia media y la desviación estándar para construir un límite variable:

$$P(T_i, m) > \overline{P(T_i)} + t_2 \cdot \sigma(T_i) \quad (26)$$

$P(T_i, m)$  es la presencia de la temática  $T_i$  durante el mes  $m$

$t_2$  es el parámetro de corrección que permitirá ajustar la definición de la señal para optimizar sus resultados. Aquí será el número de desviaciones estándares que se toman para construir el límite.

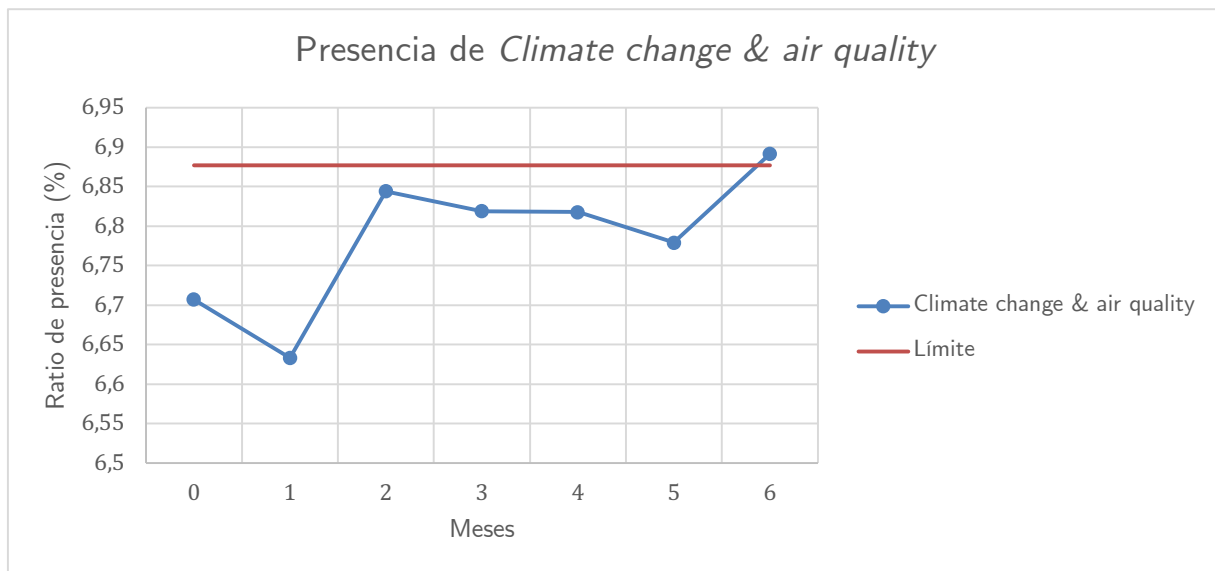


Ilustración 62 - Evolución temporal de la presencia de la temática Climate change & air quality con el límite calculado en función de los valores propios de la evolución ( $t_2 = 1$ ). (Fuente: elaboración propia)

La Ilustración 62 presenta en un caso concreto dónde se ubicaría el límite definido en la ecuación (26), en este caso, el valor en el último mes (número 6) se ubica por encima del límite, con lo cual se activará la señal (además se sabe que la temática incrementará su presencia 3 meses después). Esto entrará en el recuento de predicciones correctas de la alternativa, para aumentar su poder predictivo.

### Alternativa 3 – Ratio (diferencia log)

Si se comparan la distribución de las presencias de temáticas para las diferentes fuentes, ordenando las temáticas de más a menos importante se puede observar que las fuentes tienen una diferencia fundamental anclado en su naturaleza.

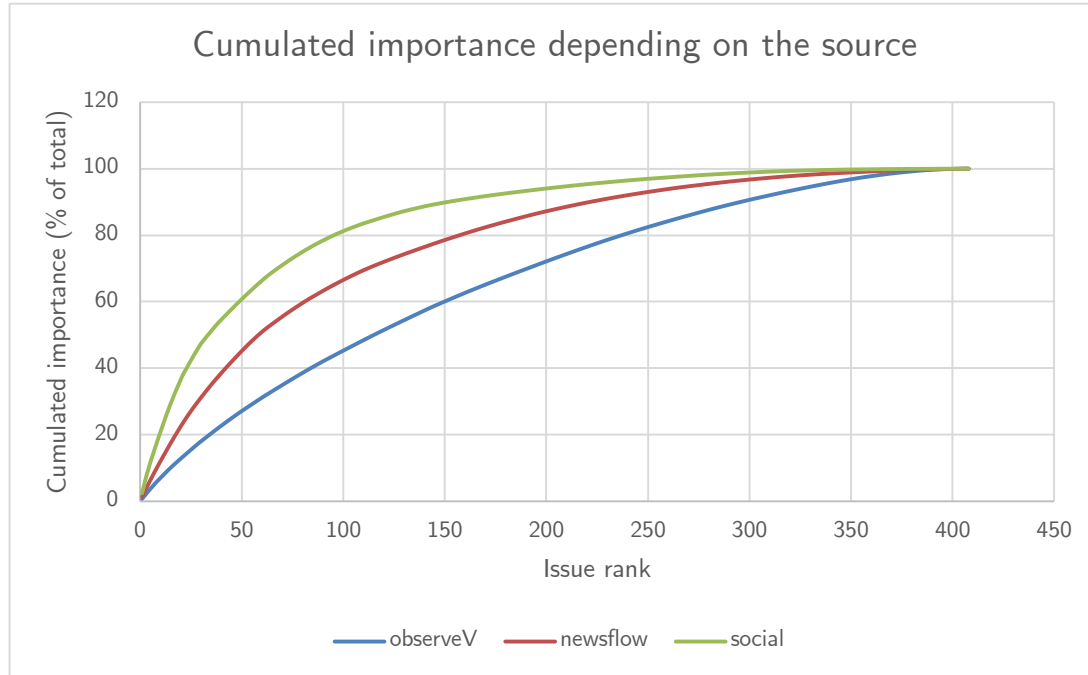


Ilustración 63 - Curva representativa de la importancia acumulada con el ranking de la temática para diferentes fuentes. (Fuente: elaboración propia)

El gráfico de la Ilustración 63 muestra las diferencias intrínsecas de comportamiento de las temáticas dentro de cada fuente. Se puede mencionar que las fuentes restantes (*benchmark*, leyes) se comportan aproximadamente como la curva de recomendaciones (*observeV*<sup>22</sup>). En el caso de la fuente *social*, y en otra medida, de *newsflow* muestra una distribución parecida a lo que se llama la ley de Pareto<sup>23</sup> en logística. En el área de los idiomas, se suele referir a la ley de Zipf, y las distribuciones de Mandelbrot. No se puede afirmar con confianza cualquier vínculo entre el comportamiento de las fuentes y esas distribuciones, pero una investigación más completa podría realmente estudiar los orígenes de aquella observación.

Más concretamente, se observa una diferencia entre *newsflow* y *social* por un lado y las otras fuentes por otro lado.

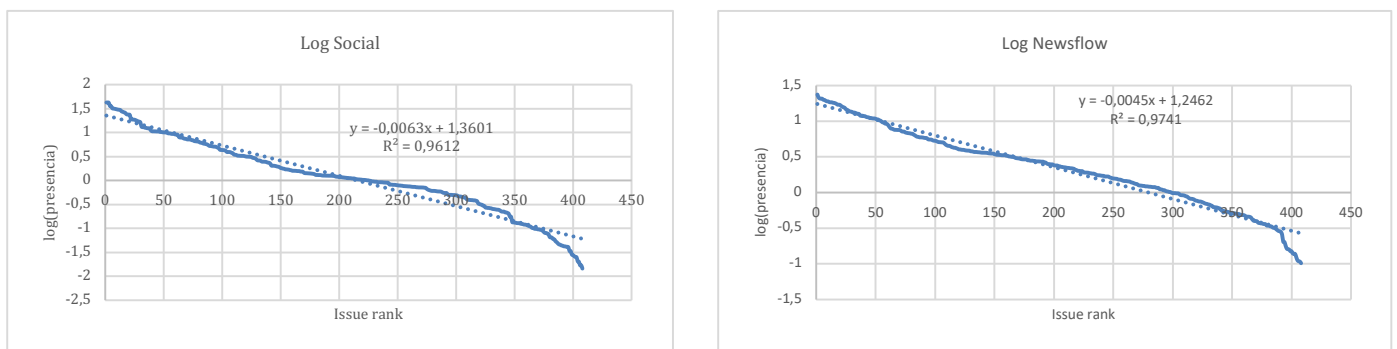


Ilustración 64 - Gráficos representando el log(presencia) para cada temática del conjunto de datos según su ranking en el conjunto total. (Fuente: elaboración propia)

La Ilustración 64 muestra que el logaritmo de la presencia para esas fuentes presenta una relación casi lineal con el ranking de la temática. Eso sirve para entender cómo se comportan las temáticas cuando ganan en importancia, el simple hecho de ganar una posición en esta clasificación

<sup>22</sup> Nombre oficial en Datamaran de la Fuente: “recomendaciones legales”

<sup>23</sup> El 20% de las causas provoca un 80% de efecto y recíprocamente

no significa incrementar su presencia de una cierta cantidad, sino multiplicarla por una cantidad. Con lo cual, parece que tendría sentido para estas fuentes, definir una señal, o test que compara la ratio de presencia entre 2 momentos a un valor dado. Se define la nueva alternativa por la condición:

$$\left\{ \begin{array}{l} \text{Si } P(T_i, m - 6) > 0, \frac{P(T_i, m)}{P(T_i, m - 6)} > t_3 \\ \text{Si } P(T_i, m - 6) = 0, P(T_i, m) > t_3 \end{array} \right\} \quad (27)$$

$P(T_i, m)$  es la presencia de la temática  $T_i$  durante el mes  $m$

$t_3$  es el parámetro de corrección que permitirá ajustar la definición de la señal para optimizar sus resultados. Aquí será el límite que la ratio tiene que superar para activar la temática.

#### Alternativa 4 – Suavizado exponencial simple

Dentro de los métodos usuales de *forecasting*<sup>24</sup> para las evoluciones temporales, se suele usar como aproximación también el método exponencial. Básicamente genera una estimación del valor futuro usando los datos anteriores, pero con una ponderación cada vez menor para los datos más antiguos. Una fórmula simple usando solo un parámetro consiste en aplicar un desarrollo límite de la función exponencial y usarla como pesos para los diferentes puntos históricos. En el caso de los 6 meses de análisis se puede calcular y activar una temática solamente si el valor predicho supera el último valor de la serie:

$$t_4 \sum_{k=0}^6 (1 - t_4)^k \cdot P(T_i, m - k) > P(m) \quad (28)$$

$P(T_i, m)$  es la presencia de la temática  $T_i$  durante el mes  $m$

$t_4$  es el parámetro de corrección que permitirá ajustar la definición de la señal para optimizar sus resultados. Aquí será el parámetro suavizante del desarrollo exponencial, toma sus valores entre 0 y 1.

### 5.4.3. Selección

Para cada una de las 3 fuentes que se van a usar para generar la señal, se seleccionará la alternativa (optimizando su parámetro primero) que permite obtener los mejores resultados dados los criterios fijados en el apartado 5.4.1. Cada alternativa solo tiene un solo parámetro que optimizar con lo cual una simple iteración sobre los valores posibles permite optimizar los resultados.

Tabla 24 - Resultados de la optimización de los parámetros para cada alternativa y cada tipo de fuente. (Fuente: elaboración propia)

Parámetros	observeV	newsflow	social
t1	0.12%	4%	0%
t2	1.3	1.2	-0.4
t3	1.15	0.95	0.93
t4	0.33	0.22	0.21

La Tabla 24 presenta los resultados de la optimización de los parámetros iterando simplemente las soluciones y seleccionando el parámetro cada vez que la solución encontrada es mejor que la mejor ya encontrada. A partir de ahora se usarán esos parámetros en las definiciones de las

<sup>24</sup> Modelos de previsiones

alternativas.

También se evaluará la exactitud de la predicción de cada solución:

$$e(S_i) = \frac{N_{correctas}}{N_{total\ temáticas}} (\%) \quad (29)$$

$N_{correctas}$  es el número de predicciones correctas (temáticas anunciadas que incrementan más temáticas no anunciadas que no incrementan)

$N_{total\ temáticas}$  es el número total de temáticas del estudio

$S_i$  es la solución número  $i$  dentro del conjunto de alternativas definidas

Esta cantidad indica el desempeño global de la solución, no sirve para evaluar la calidad de la alternativa porque una parte de la predicción es más relevante que la otra, pero permite decidir si se acepta o se rechaza una alternativa, básicamente si  $e(S_i) < 50\%$ , la alternativa no es mejor que predecir aleatoriamente, con lo cual no se aceptará.

### Recomendaciones legales

Primero es importante evaluar el desempeño de las soluciones triviales que son:

- Anunciar todas las temáticas
- No anunciar ninguno
- Anunciar aleatoriamente las temáticas (fijar  $P(A) = 0.5$ )

Tabla 25 - Desempeño de soluciones triviales con la previsión de tendencias sobre recomendaciones legales. (Fuente: elaboración propia)

observeV		Anunciar todos		No anunciar ninguno		Anunciar aleatoriamente	
Criterios	Peso	Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	0%	0%	54%	27%	54%	27%
Poder predictivo	19%	46%	9%	0%	0%	46%	9%
Selectividad	19%	29%	5%	100%	19%	50%	9%
Coherencia	9%	0%	0%	0%	0%	0%	0%
Económico	4%	100%	4%	100%	4%	100%	4%
Total	100%		18%		49%		48%

La Tabla 25 presenta los resultados obtenidos por esas 3 soluciones, esos valores constituirán un mínimo que se tiene que alcanzar por cualquier alternativa evaluada.

Los resultados de las diferentes alternativas se presentan en la Tabla 27, la alternativa que obtiene los mejores resultados será la alternativa 1: comparar la diferencia absoluta en presencia con el parámetro límite que se ha optimizado durante el proceso.

Tabla 26 - Resultados de exactitud de las diferentes alternativas. (Fuente: elaboración propia)

observeV	Alternativa 1	Alternativa 2	Alternativa 3	Alternativa 4
Exactitud	60%	55%	59%	53%

En la Tabla 26 se ve que todas las soluciones dan resultados globalmente mejores que una selección aleatoria, entonces hasta tener los resultados detallados, no se puede descartar ninguna de ellas.

Tabla 27 - Resultados de las diferentes alternativas sobre el conjunto de datos de recomendaciones legales.

(Fuente: elaboración propia)

Criterios	Peso	Alternativa 1		Alternativa 2		Alternativa 3		Alternativa 4	
		Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	58%	29%	56%	27%	56%	28%	52%	25%
Selectividad	19%	73%	14%	55%	10%	76%	14%	30%	6%
Poder predictivo	19%	90%	17%	86%	16%	93%	18%	81%	15%
Coherencia	9%	100%	9%	100%	9%	60%	6%	100%	9%
Económico	4%	100%	4%	80%	3%	100%	4%	80%	3%
Total	100%		72%		66%		69%		59%

Finalmente, la Tabla 27 presenta los resultados detallados de cada temática, las puntuaciones de coherencia y coste se ha atribuido según las justificaciones siguientes:

- Alternativa 1: la solución es sencilla de implementar y coherente con lo que se estudió: la fuente de recomendaciones legales **no** se comporta como *social* o noticias, tendría sentido que una caracterización con límite fijo sobre la variación absoluta en presencia tenga buenos resultados. Por tanto, obtiene la puntuación máxima en ambos criterios.
- Alternativa 2: basarse en datos históricos, comprimirlos en una media y desviación típica para usarlos y evaluar si la fluctuación en el último mes es normal o no parece muy coherente y sensato. No obstante, calcular una desviación típica cuesta un poco más en tiempo de cálculo que el cálculo sencillo de la primera solución, por tanto, el criterio económico no tendrá la máxima puntuación.
- Alternativa 3: calcular y evaluar la ratio de 2 presencias no tiene sentido para esta fuente porque la distribución de temáticas no lo sugiere, por tanto, no tendrá una buena puntuación de coherencia. En cuanto a coste, el simple cálculo de la ratio no cuesta mucho tiempo ni memoria y se puede considerar igual al coste de la primera alternativa.
- Alternativa 4: la solución de suavizado exponencial simple es coherente y sensata para el cliente ya que es una técnica común en previsiones simples. No obstante, es una de las más complejas desde el punto de vista de la complejidad de los cálculos.

La conclusión de la experimentación indica que la mejor alternativa sería la primera, usando la diferencia absoluta en presencia entre los meses  $m$  y  $m - 6$ . Además, se puede realizar un estudio estadístico sobre cómo esta condición clasifica correctamente las temáticas.

Se puede realizar un recuento de las temáticas agrupando según 2 criterios:

- Si pasan la prueba de la definición de la alternativa (si se activan con esa definición)
- Si realmente la presencia incrementará 3 meses después

Los resultados se ponen en la Tabla 28, las columnas presentan los recuentos según el resultado del *test* de la condición, y las filas el recuento según si la temática ganará en importancia o no.

Tabla 28 - Datos del análisis de desempeño de la alternativa 1 sobre la Fuente: recomendaciones legales.  
(Fuente: elaboración propia)

COUNT	positive test	negative test	TOTAL
increase	62	122	184
not increase	43	182	225
TOTAL	105	304	409

Se plantean las hipótesis:

- $H_0$ : La clasificación obtenida con la condición de la alternativa se distribuye de la misma manera que a nivel global (46% de incrementos)
- $H_1$ : La clasificación obtenida cambia la distribución de incrementos/no incrementos.

Se rechazará la hipótesis nula con nivel de significancia  $\alpha = 10\%$ .

Respetando la distribución de temáticas que incrementan comparado con el número total de temáticas (46% de incrementos en general), el recuento esperado se presenta en la Tabla 29.

Tabla 29 - Resultados esperados si el método de clasificación no cambiaría las distribuciones. (Fuente: elaboración propia)

COUNT	positive test	negative test	TOTAL
increase	47.2	136.8	184.0
not increase	57.8	167.2	225.0
TOTAL	105.0	304.0	409.0
p-value	0.38%	8.88%	

Dadas la distribución esperada descrita en la Tabla 29, se puede realizar una prueba del  $\chi^2$  para validar o rechazar la hipótesis nula. Los resultados se dan en la fila “p-value” de la Tabla 29, en ambos casos (para la clasificación de los positivos y de los negativos) la p-value es inferior al nivel de significancia fijado. Por tanto, se tiene que rechazar la hipótesis nula en ambos casos y se debe aceptar que el método de clasificación cambia las distribuciones de las temáticas. Además, se observa que el recuento es más grande en los casos válidos (números en negrita en la Tabla 28). Por tanto, se comprueba que los resultados obtenidos no se deben a la suerte sola, sino que la definición de la solución tiene algún poder para predecir un incremento futuro.

### Artículos de prensa

Se analizan algunas soluciones triviales para empezar, los resultados se encuentran en la Tabla 30, esos son los niveles de desempeño que se tienen que cumplir como mínimo.

Tabla 30 - Resultados de las diferentes alternativas sobre el conjunto de datos de artículos de prensa. (Fuente: elaboración propia)

newsflow		Anunciar todos		No anunciar ninguno		Anunciar aleatoriamente	
Criterios	Peso	Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	0%	0%	44%	22%	44%	22%
Poder predictivo	19%	56%	11%	0%	0%	56%	10%
Selectividad	19%	0%	0%	100%	19%	44%	8%
Coherencia	9%	0%	0%	0%	0%	0%	0%
Económico	4%	100%	4%	100%	4%	100%	4%
Total	100%		14%		44%		44%



Tabla 31 - Resultados de exactitud para las diferentes alternativas y para la Fuente: artículos de prensa. (Fuente: elaboración propia)

newsflow	Alternativa 1	Alternativa 2	Alternativa 3	Alternativa 4
Exactitud	41%	44%	47%	56%

Los resultados de exactitud para cada alternativa estudiada se presentan en la Tabla 31, se puede observar que solo una alternativa supera el mínimo de 50% exigido. Para estudiar alternativas que merecen ser comparadas, es necesario modificar las 3 primeras para que obtengan una exactitud como mínimo superior al 50%. Se pueden modificar invirtiendo la definición del test cambiando el  $>$  por un  $<$  y volviendo a optimizar el parámetro:

Alternativa 1\*<sup>25</sup>:

$$P(T_i, m) - P(T_i, m - 6) > t_1^* \quad (30)$$

Alternativa 2\*:

$$P(T_i, m) < \overline{P(T_i)} + t_2^* \cdot \sigma(T_i) \quad (31)$$

Alternativa 3\*:

$$\left\{ \begin{array}{l} \text{Si } P(T_i, m - 6) > 0, \frac{P(T_i, m)}{P(T_i, m - 6)} < t_3^* \\ \text{Si } P(T_i, m - 6) = 0, P(T_i, m) < t_3^* \end{array} \right\} \quad (32)$$

Tabla 32 - Resultados obtenidos con las alternativas modificadas. (Fuente: elaboración propia)

Alternativa	Parámetro	Exactitud
1*	0.1%	51.50%
2*	0.5	50.74%
3*	0.89	50.12%

En la Tabla 32 se muestran los nuevos parámetros y primeros resultados de exactitud para poder aceptar las alternativas en las comparaciones. Ahora se aceptan todas, aunque los resultados no sean impresionantes a nivel global.

Tabla 33 - Resultados obtenidos para las diferentes alternativas seleccionadas para la Fuente: artículos de prensa. (Fuente: elaboración propia)

Criterios	Peso	Alternativa 1*		Alternativa 2*		Alternativa 3*		Alternativa 4	
		Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	46%	23%	47%	23%	46%	23%	50%	25%
Poder predictivo	19%	57%	11%	61%	11%	60%	11%	68%	13%
Selectividad	19%	51%	10%	71%	13%	70%	13%	68%	13%
Coherencia	9%	40%	4%	60%	6%	80%	8%	80%	8%
Económico	4%	100%	4%	80%	3%	100%	4%	80%	3%
<b>Total</b>	<b>100%</b>		<b>50%</b>		<b>56%</b>		<b>58%</b>		<b>61%</b>

La Tabla 33 presenta los resultados obtenidos, las filas de coherencia y de criterio económico se justifican por lo siguiente:

- Alternativa 1\*: es la menos coherente por 2 razones principales: primero es una

<sup>25</sup> Las alternativas modificadas se marcarán con una \*

alternativa invertida (se considera que las temáticas que disminuyeron su presencia son más probables de incrementar luego, se basa en un fenómeno de regresión a la media<sup>26</sup> lo que podría significar que la clasificación no vale más que una decisión aleatoria centrada). Segundo, se introdujo la alternativa 3 (con las ratios) justamente porque *news* tenía una forma de logaritmo, en oposición a las distribuciones lineales para las cuales la alternativa 1 corresponde más. Su coste es mínimo, por su facilidad de cálculo, sobre todo.

- Alternativa 2\*: recibe una puntuación de coherencia media por ser una alternativa invertida y también porque no es la alternativa 3 que sería la más coherente. Su coste es mayor debido a las calculaciones de desviación estándar y promedio sobre los 6 meses necesarios.
- Alternativa 3\*: no recibe la máxima puntuación en coherencia porque es invertida. Su coste es mínimo.
- Alternativa 4: no es la alternativa 3 que sería la más coherente, pero tiene bastante sentido seleccionar esta. Su coste no es el mínimo para el desarrollo exponencial que se tiene que calcular.

La conclusión de la Tabla 33 es que se selecciona la solución de la alternativa 4, con el método de suavizado exponencial simple que obtiene los mejores resultados. Además, una prueba estadística del  $\chi^2$  obtiene los resultados presentados en la Tabla 34 y Tabla 35.

Tabla 34 - Recuento de temáticas según si esta actividad o no, y si incrementa o no, para el test de la alternativa 4. (Fuente: elaboración propia)

COUNT	positive test	negative test	TOTAL
increase	<b>88</b>	139	227
not increase	41	<b>140</b>	181
TOTAL	129	279	408

Tabla 35 - Resultados esperados en caso de validación de la hipótesis  $H_0$ , y *p-value* obtenida con los resultados reales. (Fuente: elaboración propia)

COUNT	positive test	negative test	TOTAL
increase	71.8	155.2	227.0
not increase	57.2	123.8	181.0
TOTAL	129.0	279.0	408.0
<i>p-value</i>	0.40%	5.05%	

La *p-value* máxima obtenida en los dos tests realizados es inferior al nivel de significancia  $\alpha = 10\%$ , por tanto, se rechaza la hipótesis  $H_0$ . Como los números en negrita en la Tabla 34 son mayores que los esperados se concluye que el test es adecuado para separar las temáticas que incrementan de las que no.

### Social media

Las opciones triviales que servirán de mínimo de comparación para el *score* total se presentan en la Tabla 36.

<sup>26</sup> Regresión a la media: cuando un evento es normalmente distribuido es necesario realizar muchos experimentos para

Tabla 36 - Resultados obtenidos por 3 soluciones triviales para la fuente de datos: social media. (Fuente: elaboración propia)

social		Anunciar todos		No anunciar ninguno		Anunciar aleatoriamente	
Criterios	Peso	Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	0%	0%	54%	27%	54%	27%
Poder predictivo	19%	46%	9%	0%	0%	46%	9%
Selectividad	19%	0%	0%	100%	19%	50%	9%
Coherencia	9%	0%	0%	0%	0%	0%	0%
Económico	4%	100%	4%	100%	4%	100%	4%
Total	100%		12%		49%		48%

Se presentan en la Tabla 37 las ratios de exactitud obtenidas por cada una de las soluciones, se tiene que descartar la alternativa 4 por no superar el mínimo del 50% requerido.

Tabla 37 - Resultados de exactitud para las diferentes alternativas y para la Fuente: social media. (Fuente: elaboración propia)

social	Alternativa 1	Alternativa 2	Alternativa 3	Alternativa 4
Exactitud	72%	75%	77%	45%

No es necesario modificar la alternativa 4 porque ya existen 3 alternativas que se podrán comparar entre sí, la 4 se puede descartar directamente.

Tabla 38 - Resultados obtenidos por las diferentes alternativas en la Fuente: social media. (Fuente: elaboración propia)

social		Alternativa 1		Alternativa 2		Alternativa 3	
Criterios	Peso	Puntuación	Score ponderado	Puntuación	Score ponderado	Puntuación	Score ponderado
Fiabilidad	49%	68%	34%	72%	36%	79%	39%
Poder predictivo	19%	81%	15%	80%	15%	74%	14%
Selectividad	19%	71%	13%	66%	12%	53%	10%
Coherencia	9%	60%	6%	80%	8%	100%	9%
Económico	4%	100%	4%	80%	3%	100%	4%
Total	100%		72%		74%		76%

La Tabla 38 muestra los resultados obtenidos, los criterios de coherencia y económico se rellenan así:

- Alternativa 1: social es la fuente que más se parece a un logaritmo, no tendría mucho sentido aplicarle una comparación simple de la diferencia en valores absolutos. No obstante, es una solución muy barata de implementar.
- Alternativa 2: no es la más coherente, pero tiene sentido adaptar el límite que hay que superar con la desviación típica y la media de su evolución histórica. Es una solución más cara en cálculos.
- Alternativa 3: es una solución muy barata en cálculos y muy adecuada con la forma de la curva de *social*, obtiene la máxima puntuación en ambos criterios.

Finalmente se selecciona la solución de la alternativa 3 (comparar la ratio de incremento con

un límite fijo) que obtiene los mejores resultados. La Tabla 39 y la Tabla 40 muestran los resultados de un análisis estadístico y un test del  $\chi^2$  usando el método de la alternativa 4 sobre el conjunto de datos. La probabilidad de obtener estos resultados considerando la hipótesis  $H_0$  verdad sería como máximo de  $1 \cdot 10^{-13}$ , lo cuál es significativamente inferior al nivel de significancia de  $\alpha = 10\%$ . Se tiene que rechazar la hipótesis  $H_0$  y como los números en negrita en la Tabla 39 son mayores de lo esperado se puede concluir que la alternativa 4 es un buen criterio de clasificación de las temáticas.

Tabla 39 - Resultados obtenidos con la clasificación de la alternativa 3 para la Fuente: social media. (Fuente: elaboración propia)

COUNT	positive test	negative test	TOTAL
increase	<b>144</b>	47	191
not increase	50	<b>175</b>	225
TOTAL	194	222	416

Tabla 40 - Resultados esperados si la hipótesis nula fuese verdad. (Fuente: elaboración propia)

COUNT	positive test	negative test	TOTAL
increase	89.1	101.9	<b>191.0</b>
not increase	104.9	120.1	<b>225.0</b>
TOTAL	<b>194.0</b>	<b>222.0</b>	<b>416.0</b>
p-value	<b>2.55E-15</b>	<b>1.00E-13</b>	

Entre todo lo que se ha estudiado para la predicción de las futuras tendencias, la alternativa que obtuvo los mejores resultados fue la Alternativa 3 para la fuente *social media* – predicción basada en la ratio de incremento en los últimos 6 meses – con un *score* de 76%.

Por fin se han seleccionado diferentes alternativas para identificar a las tendencias en las 3 fuentes dinámicas de Datamaran: recomendaciones legales, artículos de prensa y *Twitter*.

Tabla 41 - Resumen de las diferentes soluciones seleccionadas. (Fuente: elaboración propia)

Fuente	Alternativa #	Nombre	Parámetro	Ecuación
observeV	1	Sencilla	0.12%	(25)
news	4	Suavizado	0.216	(28)
		exponencial		
social	3	Ratio	0.93	(27)

La Tabla 41 describe las decisiones que se han tomado en este apartado acerca de los métodos de previsión de las tendencias, esa información se transmitirá al usuario a través de una hoja de cálculo Excel automáticamente generada en Python como presentada en la Ilustración 65.

[INCREASE IN SOCIETAL SOURCES (6 MONTHS)] - EXPLANATION				
This analytical lens should complement your annual or bi-annual materiality review.				
It assesses the change in the importance of your issues over 6 months across 3 leading sources.				
If an issue's importance has grown during that period and according to a source, the cell value is TRUE.				
Issues with a 'Total' value of 3 present stronger emergence; 2 is weaker; 1 and 0 are relatively stable.				
SOCIETAL SIGNALS TO CONSIDER		INCREASE IN SOCIETAL SOURCES (6 MONTHS)		
Issue name	Social	Newsflow	Voluntary regulation	Total
Biodiversity	FALSE	FALSE	FALSE	0
Business Ethics	FALSE	FALSE	FALSE	0
Climate & Energy	FALSE	FALSE	FALSE	0
Digitalization & Data Security	FALSE	FALSE	TRUE	1
Disclosure	TRUE	TRUE	FALSE	2
Customer Engagemant	FALSE	FALSE	FALSE	0
Emissions (Air & Soil)	TRUE	FALSE	FALSE	1
Employee Engagement & Diversity	FALSE	FALSE	FALSE	0
Geopolitical Risk	TRUE	TRUE	TRUE	3
Governance Structure	FALSE	FALSE	FALSE	0
Health & Safety/Product Stewardship	TRUE	FALSE	FALSE	1
Human Rights	FALSE	FALSE	FALSE	0
Innovation	FALSE	FALSE	FALSE	0
Organizational Resilience	FALSE	FALSE	FALSE	0
Responsible Supply Chain	TRUE	FALSE	FALSE	1
Stakeholder Engagement	FALSE	TRUE	FALSE	1
Sustainable Finance	FALSE	FALSE	FALSE	0
Ressource Efficiency & Waste	FALSE	FALSE	FALSE	0
Water	FALSE	FALSE	FALSE	0
© Datamaran. All rights reserved - www.datamaran.com				
Summary		Emerging Issues Analysis		

Ilustración 65 - Ejemplo de hoja de cálculo conteniendo las previsiones del modelo seleccionado para las tendencias. (Fuente: elaboración propia)

#### 5.4.4. Crítica de la selección

En la mayoría de los casos, el desempeño de las alternativas, incluso de la mejor, se ubicaba entre 50% y 75%, lo mismo para la exactitud global de la previsión. En general para hablar de una precisión aceptable se pide como mínimo el 80% - 90% de fiabilidad global de la predicción. Obviamente, en este caso el tema es bastante complicado, las temáticas no siguen un mecanismo perfectamente identificado o solo caótico, no se sabe describir lo que pasa realmente en internet y en la sociedad en general. Es importante recordar que el valor añadido para el usuario es sobre todo poder descartar temáticas menos importantes de las que son de verdad. Sobre todo, lo que vende Datamaran es una herramienta de análisis, basado en datos reales pero evaluados según sus criterios, y no un modelo de predicción (aunque se diseña como si fuese uno). En general, para este estudio, faltarían mucho más datos para confirmar las observaciones que se hicieron (incluso las que son estadísticamente significativas). También se necesita desarrollar más variantes y alternativas para clasificar, por ejemplo, si se tuviera mucho más datos, se podría entrenar un algoritmo de Deep learning que trabajaría sobre los diferentes datos para intentar extraer una lógica que permita clasificar correctamente las temáticas antes de que incrementaran. Posiblemente se necesitan incluir más informaciones que la presencia, pero también el ranking de las temáticas en las diferentes fuentes, por ejemplo.

La elección del horizonte temporal de 3 meses para identificar las tendencias que se confirman viene más de una restricción por la cantidad de datos que se puede encontrar que por una justificación desde el punto de vista del usuario. Se podría realizar encuestas sobre qué horizonte temporal las tendencias interesan a los clientes para adaptar el estudio a sus necesidades.

La determinación de los pesos en la matriz de jerarquías también podría ser objeto de una encuesta más profunda dentro del equipo responsable de las comunicaciones con los clientes para identificar más precisamente las necesidades y los objetivos del producto.



No obstante, la previsión de la tendencia para la fuente “social” es bastante interesante y significativamente más precisa que para las otras, lo que parece indicar de manera contra intuitiva que Twitter es más predecible que los textos de recomendaciones redactados por las ONGs de una cierta manera.

## 5.5. Conclusión

El objetivo de este capítulo era diseñar lo más precisamente posible la solución más adecuada para resolver los requerimientos R11 y R12. Las soluciones serán un conjunto de condiciones simples que permiten lógicamente determinar si una temática presenta las características adecuadas para ser importante o no, y así ayudar al usuario focalizar su energía sobre las más importantes.

Los responsables del producto requieren que se den recomendaciones de manera estática (usando el estado del análisis en un momento dado) y de manera histórica (analizar la evolución temporal de los datos). Por ello se crearon dos hojas Excel que contendrán las previsiones calculadas gracias a 4 operaciones lógicas sencillas usando los datos reales del análisis de materialidad estudiado.

En general el estudio demuestra que la previsión es una disciplina compleja, para la cual las operaciones sencillas no dan muy buenos resultados, pero gracias a una amplia propuesta de soluciones diferentes se ha podido seleccionar algunas que demuestran su desempeño sobre un conjunto de análisis reales y de manera estadísticamente significativa.

Se podría mejorar el análisis primero proponiendo más alternativas, o generando nuevas gracias a la inteligencia artificial y el Big Data, pero también se podría cuestionar el entorno global del estudio (los meta parámetros que se podrían modificar como el horizonte temporal y las informaciones que se usan para construir el modelo).

## 6. Implantación de la herramienta

### 6.1. Introducción

En los capítulos anteriores se presentaron el servicio de Datamaran, el proyecto de mejora del módulo de monitoring que se planteaba así que el trabajo de investigación que se hizo sobre las señales sociales que se querían detectar para las temáticas. En los requerimientos se explicitaron por una parte los requerimientos necesarios en la interfaz y arquitectura del módulo, y se dejaron libres las metodologías empleadas para la generación de indicadores con el objetivo de focalizar la atención del usuario sobre algunas temáticas importantes.

En este capítulo se detalla el proceso de implantación de la herramienta en la plataforma Datamaran. Se presentarán la planificación de las tareas necesarias, las medidas tomadas para gestionar los fallos potenciales y se evaluará el interés económico de este trabajo.

Primero se explicará el plan de implantación del nuevo módulo, detallando las tareas necesarias, el cronograma y los recursos necesarios. Luego, el plan de contingencias presentará el Análisis de Modos de Fallos y sus Efectos (AMFE) de las medidas tomadas en función de la priorización de las correcciones necesarias. También se harán estimaciones de beneficios obtenidos gracias a este trabajo de manera que se pueda calcular un horizonte de retorno sobre la inversión necesaria. Finalmente, se presentará el nuevo módulo así como los nuevos visuales y archivos que lo constituyen.

### 6.2. Plan de implantación

En otoño 2020 se tenía que desplegar la nueva versión de la plataforma: Datamaran5, por la pandemia de Covid-19 se tuvo que aplazar este despliegue a principios de 2021. No obstante, desde el lanzamiento de Datamaran 4 y la primera versión del módulo de *continuous monitoring*, el equipo se planteaba un *update* necesario antes de Datamaran5. A principios de 2020 se decidió que el nuevo módulo de *monitoring* debería de presentarse al público el 22 de abril (día de la Tierra). Para esta ocasión se hará un comunicado a todos los clientes que el módulo de continuous monitoring ha sido actualizado con un nuevo diseño, y nuevas funcionalidades (posibilidad de exportar los datos, de tener un diagnóstico de las temáticas). Por tanto, esta fecha constituye una fecha límite del proyecto, para la cual el módulo tiene que estar listo, testeado y funcionando para el *Webinar* de presentación y formación de los usuarios.

La lista de los participantes y responsables del proyecto se presentan en la Tabla 42 a continuación.



Tabla 42 - Lista de responsables y participantes al proyecto de mejora de continuous monitoring. (Fuente: elaboración propia)

Personal	Rol
Responsable 1	CTO
Responsable 2	Director of Product
Responsable 3	Main Software architect
Responsable 4	Data Scientist/Back end developer
Responsable 5	Junior Data Scientist
Consultor 1	Director of Customer Success
Consultor 2	Senior data scientist
Participante 1	Creative director
Participante 2	Web designer
Participante 3	UX designer
Participante 4	Front-end developer
Participante 5	Front-end developer
Participante 6	Quality assurance
Participante 7	Quality assurance
Participante 8	Data Engineer

En este proyecto participa una gran parte del personal de Datamaran (en total la empresa cuenta con 60 empleados solo), pero cabe destacar que la mayoría de los participantes no solo trabajarán en este proyecto sino en otros en paralelo. Durante todas las fases del producto intervendrán incluso más empleados de Datamaran como consultores sobre aspectos en los cuáles son expertos (desde la arquitectura de la base de datos, hasta las expectativas de los clientes).

En la Ilustración 67 se presenta el cronograma previsto del proyecto, como en todos los proyectos, el cronograma real es diferente y los requerimientos evolucionaron. Gracias al método SCRUM, la construcción del nuevo módulo se hará por *features* que los *Product Owners* priorizarán. Por ejemplo, la creación de las señales (objeto principal del capítulo 5 de esta memoria) era muy importante para el valor de la plataforma, pero su simple existencia era más prioritaria que su justificación. Por tanto, se puede implementar una solución provisional y perfeccionarlo después del despliegue.

Tabla 43 - Recursos humanos necesarios para el proyecto

#	Tarea	Personal
1	Investigar y cuestionar el funcionamiento actual del módulo	Responsables 1-5
2	Definir los nuevos requerimientos obligatorios	Responsables 1-5 + Consultores
3	Crear el esqueleto de código back end que recibirá las funcionalidades	Responsables 4-5
4	Escribir el código front end que llamará al back end	Responsable 3 + Participantes 1-5
5	Definir una metodología para decisión	Responsable 5
6	Buscar y definir alternativas para las señales	Responsable 5
7	Presentar la solución elegida	Responsables 1-5 + Consultores
8	Modificar la solución, los visuales	Responsables 1-5 + Participantes 1-5
9	Testear	Responsables 4-5 + Participantes 6-7
10	Instalar sistema de detección de fallos	Responsables 4-5 + Participante 8
11	Desplegar	Responsables 3-5
12	Presentar y formar a los clientes	Consultor 1
13	Vigilar continuamente y tomar notas para Datamaran5	Responsables 1-5 + Participante 8

La Tabla 43 detalla para cada tarea cuáles son los empleados que han de trabajar en ella, en cuanto a recursos materiales, para todas son los mismos:

- Ordenador
- Conexión internet
- Acceso a la cuenta de Datamaran en AWS
- Acceso a los servidores de Datamaran en MongoDB

El avance del trabajo ha sido controlado usando como medida el número de páginas de esta memoria, una aproximación discutible pero que muestra por ejemplo que se tuvo que esperar al 19 de abril que todo esté bien definido para redactar la mayor parte de los detalles técnicos. La evolución de la memoria está representada en la Ilustración 66.

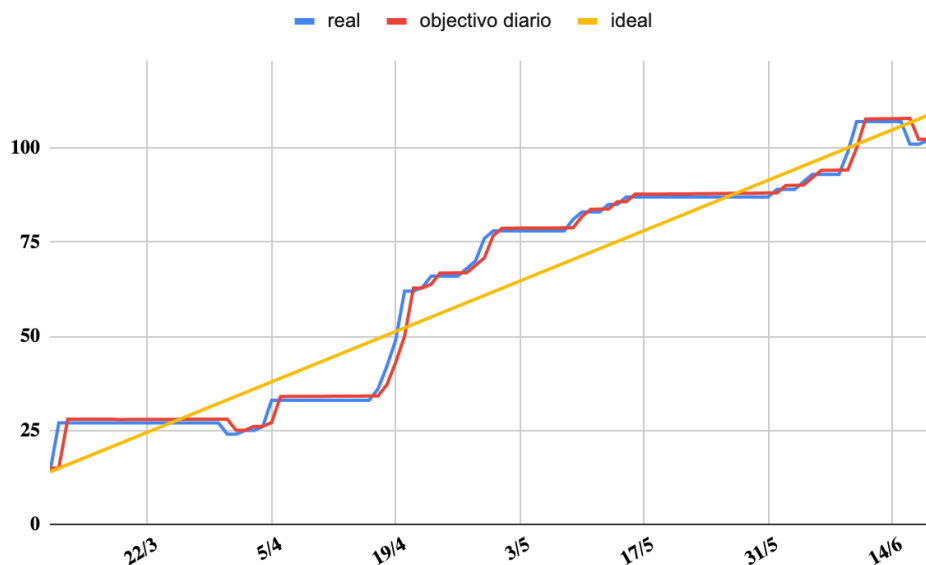


Ilustración 66 - gráfico de evolución del número de páginas de la memoria. (Fuente: elaboración propia)

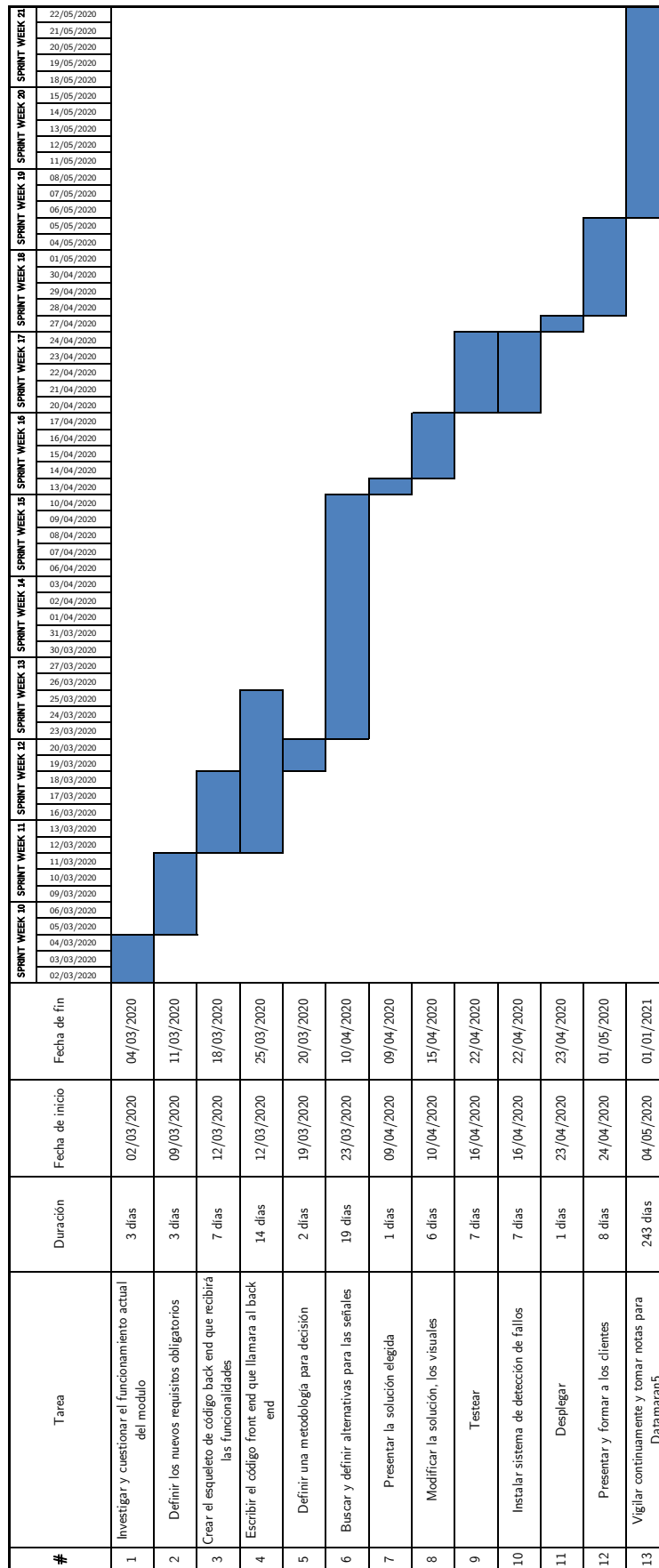


Ilustración 67 – Cronograma del proyecto de mejora del módulo de continuous monitoring (Fuente: elaboración propia)

## 6.3. Plan de contingencias

Cómo se ve en la Ilustración 67, la última tarea programada es una vigilancia continua del proyecto, en efecto, una vez se ha desplegado correctamente el software varios errores/problemas pueden ocurrir durante el ciclo de vida del módulo .

La tarea 10 que consiste en instalar un sistema de detección de fallos se realizó gracias a ciertas funcionalidades del módulo *AWS CloudWatch* de Amazon. Este módulo deja a sus usuarios programar alarmas que irán vigilando los *logs*<sup>27</sup>. Simplemente programando que el módulo envíe una notificación por *email* al equipo cuando se detecta un error para que se puedan estudiar las causas.

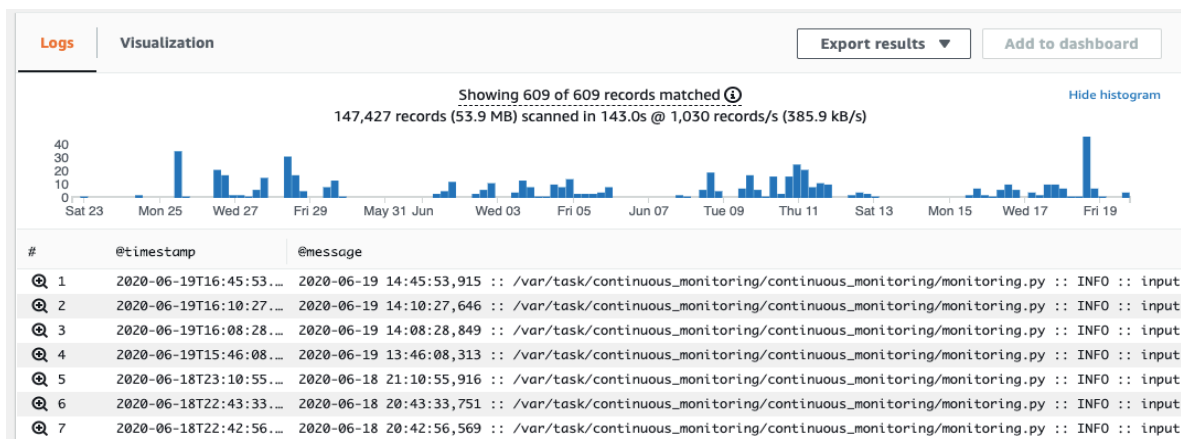


Ilustración 68 - Panel de vigilancia de la actividad en las lambdas. (Fuente: Amazon Web Service Console)

La Ilustración 68 presenta en la página de AWS<sup>28</sup> cómo se puede vigilar la actividad en la funcionalidades *Lambda* usadas en el código del módulo , así como los *logs* que describen lo que ha pasado.

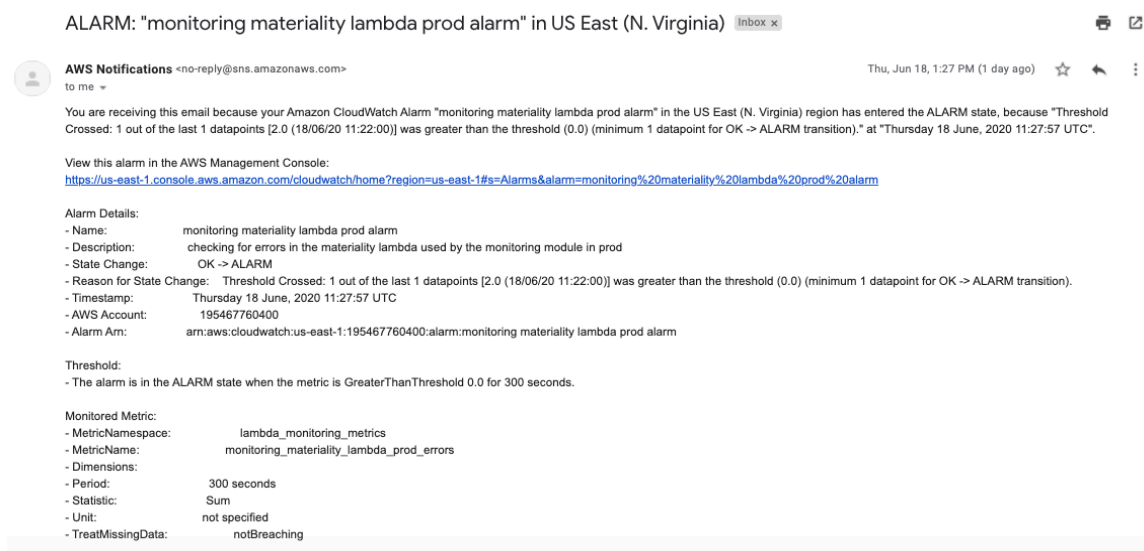


Ilustración 69 - Ejemplo de notificación por correo electrónico de detección de un problema con las lambdas de AWS. (Fuente: Gmail)

<sup>27</sup> Texto escrito por la lambda durante su ejecución, se puede escribir código que describe en tiempo real lo que está pasando en la lambda para hacer un *debug* más fácil

<sup>28</sup> Amazon WebService

En la Ilustración 69 se muestra como AWS puede notificar a los miembros del equipo cuando se lanza un error desde las lambdas.

Pero este método de vigilancia es una buena herramienta de detección de los problemas, pero no permite prevenir que los problemas ocurran. Para ello se tienen que implementar en el código una serie de instrucciones para detectar algo que está pasando mal, así como un código de test por separado que irá verificando que en condiciones dadas el resultado obtenido corresponde a lo esperado. En programación se suele hablar de *unit tests*, se refieren a algoritmos cortos que tratan de verificar que una funcionalidad del código global está dando los resultados esperados. Para ellos se suele seguir el esquema presentado a continuación en la Ilustración 70, son un proceso de calidad muy usado en programación para dos razones principales:

- Cambios e integración: cuando un equipo trabaja en el mismo archivo, se necesita tener un buen control de versiones para asegurar que nadie está borrando el trabajo de otros, los unit tests permiten verificar, antes de aceptar un cambio en el código, que el nuevo código escrito no rompa con la lógica anterior. La principal ventaja también es que esto se hace de manera automatizada de forma muy rápida e identifica bastante precisamente de dónde viene el fallo, o cuál ha sido la funcionalidad cambiada.
- *Definition of done* y *test driven development*: cada vez más se desarrolla una buena práctica a la hora de escribir código: definir un test que confirme que la funcionalidad está funcionando **antes** de empezar a escribir el código de aquella *feature*. Eso fuerza al desarrollador a pensar en su *definition of done*, es decir: ¿bajo cuales condiciones se puede aceptar un código como realizando las funciones requeridas? Si antes de empezar el código se redacta una serie de instrucciones que deben acabar correctamente **solamente si** el código efectúa las buenas operaciones se comprobó que se escribe el código de manera más eficiente. Además, el test, una vez escrito tiene que fallar al principio ya que la *feature* no está realizada, y el objetivo es conseguir que el test acabe correctamente sin modificarlo directamente.

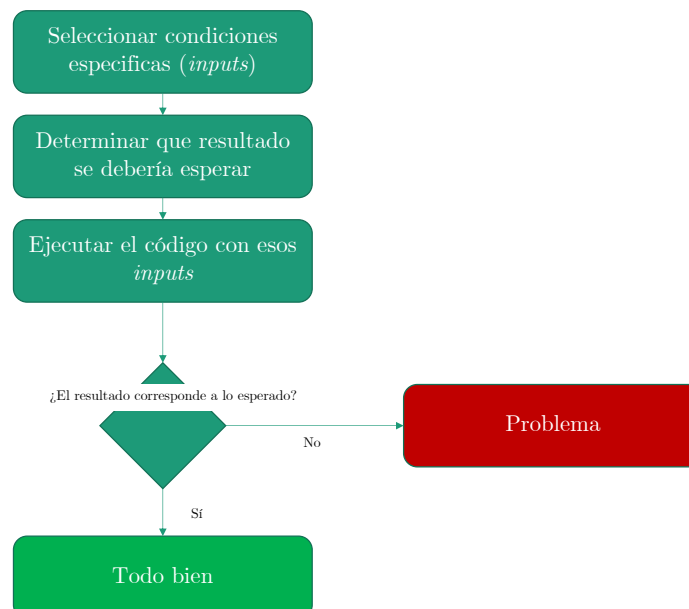


Ilustración 70 - Esquema de proceso de los unit tests. (Fuente: elaboración propia)

Tabla 44 - Análisis de Modos de Fallos y Efectos del proyecto de mejora. (Fuente: elaboración propia)

Descripción de la fase	Modo/s potencial/es de fallo	Efecto/s potencial/es del fallo	Gravedad	Causa(s) potencial(es) del fallo(s)	Ocurrencia	Verificación(es) y/o control(es) actual(es)	Detección n	NPR	Acción(es) recomendada(s)	Área(s) / persona(s) responsable(s) y fecha de realización
Descarga del archivo Excel con las señales	Todas las temáticas están activadas	El cliente esta confuso, el servicio no aporta valor	5	Error de calculo	5	Test de los casos limites	8	200	Corregir la lambda	Quality assurance y Back end
Click al botón para ver el grafico	No se muestra el grafico	El cliente se frustra	5	Error en el front end	2	Test de los casos limites	7	70	Gestionar los errores	Quality assurance y Front-end
Selección de un análisis	No se carga el buen análisis	El cliente esta confuso	5	Confusión en los id de análisis	1	Mensaje de error	10	50	Gestionar el error en front end	Quality assurance y Front end
Selección de un análisis	La pagina no carga nada	El cliente se enfada	8	El servidor no esta funcionando/ El calculo ha provocado un error en la lambda	2	Alarmas AWS	3	48	Testear todos los casos limites	Quality assurance y Back end
Descarga del archivo Excel resumiendo los cambios	No se descarga la Excel	El cliente se enfada	8	Error en la lambda	2	Alarmas AWS	3	48	Corregir la lambda	Quality assurance y Back end
Selección de una temática	Los datos de la tabla no corresponden a los del grafico	El cliente se frustra	5	Error en los cálculos de las lambdas	1	Alarmas AWS	2	10	Corregir la lambda	Quality assurance y Back end
Cargamento de los resultados	La base de datos esta en mantenimiento	El cliente se enfada	8	Falta de espacio, el cluster reinicia	1	Alarmas MongoDB/Alarmas AWS	1	8	Borrar datos inútiles, migrar la base de datos, comprar mas espacio	Data Engineer

La Tabla 44 presenta el AMFE de la implantación de la herramienta, el principal efecto que puede tener un fallo es básicamente que el cliente no esté satisfecho (a diferentes niveles). Obviamente muchos otros problemas pueden ocurrir y generar un riesgo en la infraestructura de la red, pero la mayoría de estos problemas están gestionados por proveedores de servicios como AWS y MongoDB. No se mencionan aquí varios de ellos porque el poder decisional de Datamaran acerca de su resolución es casi nulo.

Cabe destacar que uno de los riesgos impredecibles al proyecto habrá sido la pandemia de Covid-19 que hubiera podido amenazar seriamente la programación del proyecto. No obstante, la administración de Datamaran, antes de que se declarara oficialmente el encerramiento obligatorio, propuso que todos los empleados trabajaran desde casa, ya que el trabajo se podía hacer enteramente a distancia. Al ser una empresa de apenas 5 años, el riesgo para Datamaran durante la crisis era que por el impacto sobre la economía mundial tuviera un impacto directo sobre las finanzas de los clientes y que rápidamente la empresa se quedase sin ingresos suficientes para pagar a los empleados, aunque seguían trabajando. La situación obviamente es complicada y necesita que pase el tiempo para analizar correctamente el impacto de “*The Great Lockdown*” (Gopinath, 2020) sobre el portafolio de clientes de Datamaran. De forma general, la empresa supo gestionar la crisis correctamente y sufrió poco de ella.

## 6.4. Valoración económica

Por razones de confidencialidad, ciertos de los datos de este apartado se han modificado. Una vez realizado el presupuesto del trabajo, se tiene que estimar cómo se van a valorar sus resultados con el departamento financiero.

Tabla 45 - Diferentes escenarios para el retorno sobre inversión del proyecto. (Fuente: elaboración propia)

	Peor escenarios	Escenario más probable	Mejor escenario
<b>Consolidación del portafolio</b>			
Clientes convencidos	5	10	15
Responsabilidad del módulo	5%	10%	15%
Ingresos anuales adicionales atribuidos al módulo	7 787,50 €	31 150,00 €	70 087,50 €
<b>Nuevos clientes que suscriben</b>			
Clientes convencidos	0	5	5
Responsabilidad del módulo	0%	5%	10%
Ingresos anuales adicionales atribuidos al módulo	- €	7 787,50 €	15 575,00 €
<b>Total</b>			
Ingresos anuales adicionales atribuidos al módulo	7 787,50 €	38 937,50 €	85 662,50 €
Horizonte ROI <sup>29</sup> (meses)	27	12	12

La Tabla 45 muestra diferentes estimaciones para los beneficios obtenidos por este proyecto. Cabe recordar que el presupuesto asciende a un total de 17 344,00€, cómo Datamaran cobra las

<sup>29</sup> Return on Investment



suscripciones cada fin de año de la suscripción, se impone como seguridad un mínimo de 12 meses de retorno sobre inversión. Se toma como base el precio de un año de suscripción a la plataforma Datamaran; 35 000\$ es decir actualmente 31 150€ aproximadamente. Con esas cifras, se plantean los diferentes casos posibles.

### Peor escenario

En el peor caso, ningún nuevo cliente suscribe a la plataforma, pero como mínimo se sabe que algunos clientes del portafolio actual van a renovar su abono el año siguiente. Como mínimo se estima que 3 de ellos lo harán gracias al nuevo módulo por parte, como lo puede mostrar el *feedback* recibido después del *Webinar* de lanzamiento de la herramienta (Ilustración 71).

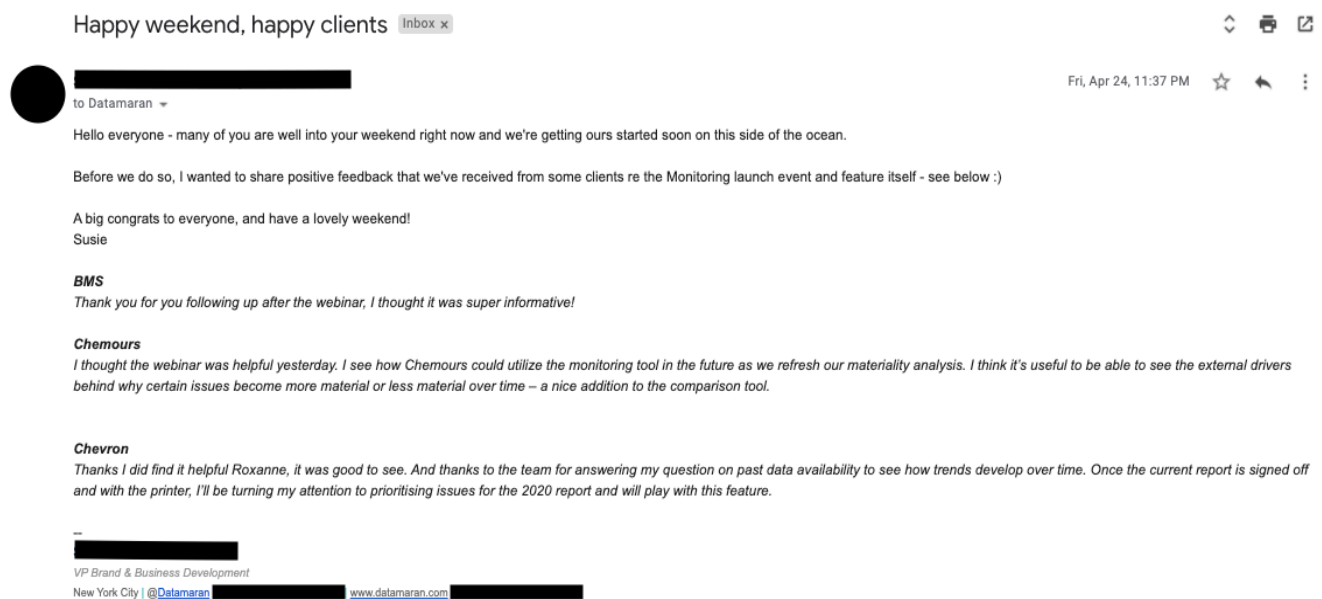


Ilustración 71 - Mail enviado por la VP Brand & Business de Datamaran después del Webinar de formación de los clientes acerca del nuevo módulo . (Fuente: Datamaran)

Como mínimo, se puede estimar que el trabajo realizado cuenta por el 5% de las razones que hacen que esos clientes se quedarán con Datamaran cuando se caduque su suscripción anual.

### Mejor escenario

En el mejor caso se puede estimar que 15 de los 60 clientes ya presentes en el portafolio renuevan su contrato al final de la suscripción gracias al nuevo módulo . En promedio, la presencia de este nuevo modelo explica en promedio 15% de las razones que los clientes tienen de renovar. Además, después de la nueva presentación video del módulo (Datamaran, 2020), así que el *Webinar* organizado puede que 5 clientes decidan suscribir un abono anual, por lo menos un 10% gracias a los resultados de este trabajo.

### Caso más probable

Las estimaciones más probables se encuentran entre los dos extremos y se detallan en las cifras de la Tabla 45. En este caso y como en el mejor escenario, el tiempo mínimo de retorno se estima a un año para tomar en consideración que no habrá flujo de caja antes del fin de la suscripción de cada uno de los clientes potenciales considerados. Por tanto, el periodo de retorno de inversión se puede estimar a un año con confianza aceptable. En el caso más probable, la tasa de rentabilidad a un año sería de un 125% del importe invertido en caso de que las estimaciones sean correctas, pero se tiene que tomar esta cifra con cuidado ya que el presupuesto del trabajo no incluye las

horas de seguimiento de los clientes, el marketing posterior a la fecha actual para convencer a los potenciales usuarios.

## 6.5. El nuevo módulo

Ahora que se han descrito los detalles de la implantación, cabe demostrar lo que se realizó en este proyecto. La Ilustración 72 presenta la nueva tabla de monitoring, y la manera de responder a los requerimientos R1 R2 R3 R4 R5 R6 y R7, cómo detallados en la Ilustración 54. En particular el botón “Show top 5” permite mostrar solamente las 5 primeras temáticas, o dejar el usuario desarrollar la lista completa de sus temáticas (requerimiento R2).

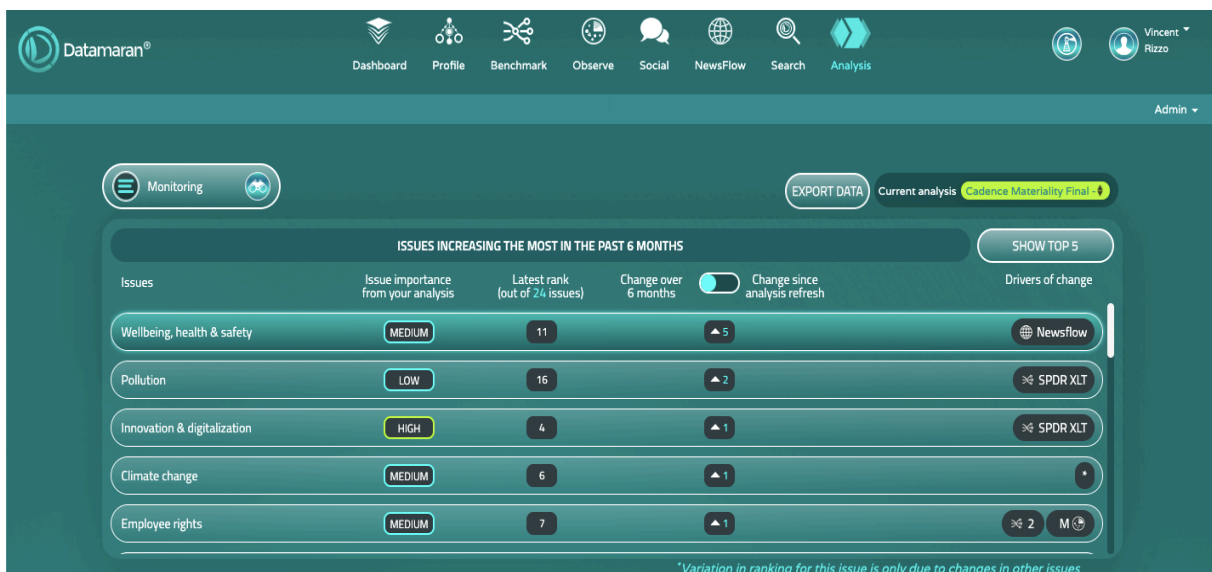


Ilustración 72 - Nueva tabla del módulo de monitoring. (Fuente: Datamaran)

De la misma manera, la Ilustración 73 presenta el nuevo gráfico y sus funcionalidades que cumplen con los requerimientos R4,R8,R9 y R10.

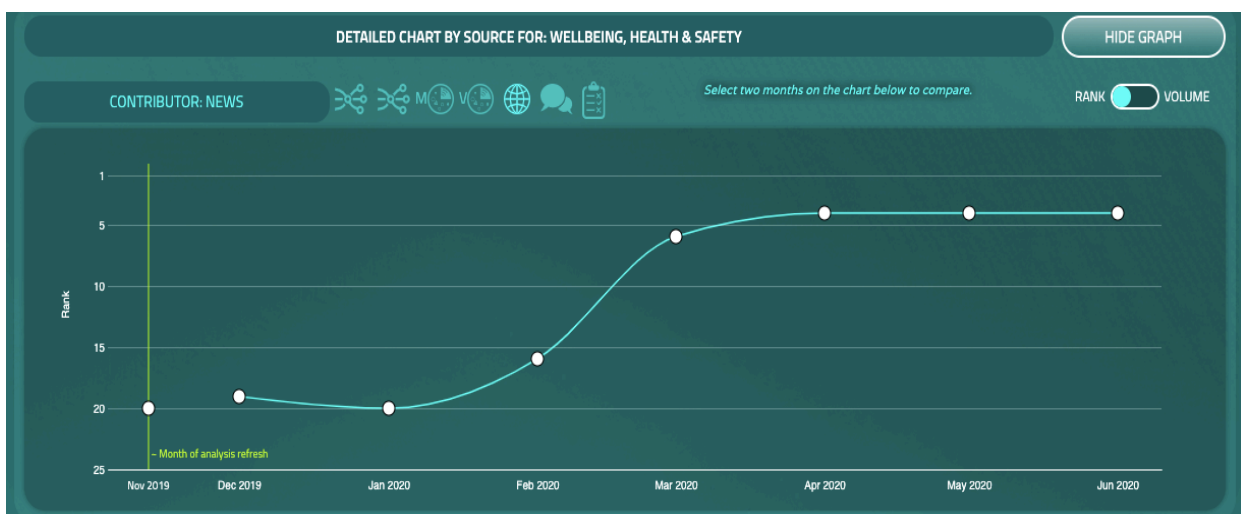


Ilustración 73 - gráfico en el nuevo módulo de continuous monitoring. (Fuente: Datamaran)

El botón disponible en la Ilustración 72, con el texto “Export data” deja al usuario descargar un archivo Excel conteniendo toda la información sobre la evolución en el ranking de las temáticas.

La Ilustración 74 presenta una primera hoja de cálculo con datos generales esenciales para que el usuario entienda lo que contiene su archivo. Entre otras cosas, ve que tendrá 2 hojas diferentes, una comparando la situación actual con la que era 6 meses antes, y otra comparándola con la situación en el mes de la ejecución más reciente de su análisis de materialidad. Este mes se identifica en la Ilustración 73 gracias a la línea amarilla en el gráfico (noviembre 2019 en el ejemplo).

	A	B
1	<b>SUMMARY</b>	
2	Analysis Name	Cadence Materiality Final -Nov 2019
3	Date of last analysis refresh	2019-11-05
4	6 months date range	2019-12 to 2020-06
5	Exported on (UTC)	2020-06-19 19:59:47
6		
7	<b>SHEETS EXPLANATION</b>	
8	6 month issue monitor	Overview of the changes in the relative ranks of your issues over the past 6 months
9	Analysis monitor	Overview of the changes in the relative ranks of your issues since your analysis was last refreshed
10		
11		
12	© Datamaran. All rights reserved - <a href="http://www.datamaran.com">www.datamaran.com</a>	
13		

Ilustración 74 - Hoja "Summary" del archivo exportable del nuevo módulo de monitoring. (Fuente: Datamaran)

En la Ilustración 75 se ve la tabla obtenida en la hoja de cálculo correspondiente a la diferencia entre el análisis hace 6 meses y como sería si se ejecutará hoy en día. Habría otra igual para la otra comparación de datos. Se puede ver que encima de cada tabla que se presenta aparecen algunas líneas de texto para recordar lo que se está observando. En efecto, la formación de los clientes ocupa una gran parte de los recursos invertidos en Datamaran, porque muchas veces se quieren presentar datos precisos y complejos pero los usuarios no logran entender las sutilidades de ello.

The below table reflects the changes in the relative ranks of your issues over the past 6 months, overall and by stakeholder source. The rank change by stakeholder source is included under each source column - positive numbers indicate an increasing rank and negative numbers a decreasing rank. The source changes are calibrated by the 'source weights' defined in your analysis and aggregated to reflect an overall change in 'stakeholder score'. The final issue ranks according to the latest month's data are presented in the last column. Rank 1 reflects the greatest importance and rank N the least importance of the N issues you have defined.

Source weights (%)	ISSUE RANK CHANGES OVER 6 MONTHS										Latest external issue rank
	6 months prior	Latest month	17.5 Δ proxypeer_benchmark	17.5 Δ SPDR XLT	17 Δ Mandatory regulation	17 Δ Voluntary regulation	10 Δ Newsflow	5 Δ Social	Δ Overall stakeholder score	Latest stakeholder issue rank	
Wellbeing, health & safety	2019-12	2020-06	-1	0	2	2	15	8	5	11	
Pollution	2019-12	2020-06	1	3	0	-1	1	0	2	16	
Innovation & digitalization	2019-12	2020-06	0	2	0	1	0	0	1	4	
Climate change	2019-12	2020-06	0	-1	0	0	-1	0	1	6	
Employee rights	2019-12	2020-06	0	1	1	0	-1	-2	1	7	
Skilled workforce	2019-12	2020-06	4	5	0	0	-4	0	1	12	
Product Stewardship	2019-12	2020-06	0	0	-1	-1	1	9	1	14	
Responsible investment	2019-12	2020-06	0	0	0	0	-1	0	1	23	
Customer privacy and information security	2019-12	2020-06	0	0	0	0	0	0	0	1	
Governance	2019-12	2020-06	0	0	0	1	-1	0	0	2	
Employee development	2019-12	2020-06	0	-1	0	-2	-4	0	0	3	
Fair and inclusive workplace	2019-12	2020-06	0	1	0	0	-2	0	0	5	
Transparency	2019-12	2020-06	-3	-1	0	-1	2	-4	0	9	
Customer responsibility	2019-12	2020-06	0	0	0	0	0	0	0	10	
Natural capital	2019-12	2020-06	1	0	-1	1	-1	-4	0	19	
Solvency and economic value	2019-12	2020-06	0	2	0	0	2	0	0	20	
Inclusion & accessibility	2019-12	2020-06	0	0	0	0	-4	1	0	21	
Employment practices	2019-12	2020-06	0	1	0	0	2	0	0	22	
Local community support	2019-12	2020-06	0	1	0	-2	4	0	-1	13	
Water	2019-12	2020-06	0	-4	0	0	0	-3	-1	18	
ESG governance structure	2019-12	2020-06	0	0	0	1	-2	-2	-1	24	
Energy efficiency	2019-12	2020-06	0	-4	0	0	0	-1	-2	8	
Human rights	2019-12	2020-06	0	-1	-1	-1	-5	0	-3	17	
Waste management	2019-12	2020-06	-1	-2	0	-2	-1	-2	-4	15	

© Datamaran. All rights reserved - www.datamaran.com

Ilustración 75 - Hoja "6 months view" del archivo exportable del nuevo módulo de monitoring. (Fuente: Datamaran)

Por lo que es del objeto principal de este trabajo: el archivo conteniendo el análisis de tendencias, se encuentra descargable desde un nuevo botón, encima de la tabla de materialidad, como visible en la Ilustración 76.

The screenshot shows a software interface with a 'Matrix table' header. It includes tabs for 'ABSOLUTE RANK' and 'PERCENTILE RANK', and buttons for 'UPDATE ANALYSIS' and 'Download emerging issues analysis' (circled in red). Below the header is a table with columns for 'Issue Name', 'external', 'internal', and 'relevance'. The table lists various issues such as 'Climate change & air quality', 'Competitive pressure', 'Customer responsibility', etc., with numerical values and relevance levels (High, Low, Medium).

Issue Name	external	internal	relevance
Climate change & air quality	1	2	High
Competitive pressure	16	17	Low
Customer responsibility	17	21	Low
Employee rights	3	8	High
Employment practices	25	23	Low
Energy efficiency	14	9	Medium
Ethics & compliance	2	18	Medium
Fair and inclusive workplace	13	17	Medium
Geopolitical events	9	17	Medium
Governance	10	10	Medium
Human rights	19	19	Low
Inclusion & accessibility	21	22	Low
Innovation & digitalization	7	1	High
Local community support	22	24	Low
Natural capital	12	14	Medium
Pollution	5	21	Low
Product responsibility	8	8	High
Responsible pricing	24	25	Low
Responsible supply chain	23	14	Low
Solvency and economic value	18	8	Medium
Talent & development	20	14	Low
Transparency	15	8	Medium
Waste management	4	14	Medium
Water	11	8	High
Wellbeing, health & safety	6	8	High

Ilustración 76 - Ubicación del nuevo botón para exportar el archivo Excel conteniendo el análisis de temáticas presentando señales y futuras tendencias. (Fuente: Datamaran)

Una vez el archivo descargado, contiene 2 hojas, una parecida a la de la Ilustración 74 conteniendo explicaciones generales, y la otra será como la de la Ilustración 77. Se observa que las señales se han juntado en una única tabla separando los dos aspectos por una columna total. El objetivo es que los usuarios puedan sortear las temáticas como quieran y usar las dos tablas como una. Por una parte, si la temática presenta signos de futuro incremento porque últimamente hubo más actividad en fuentes sociales (parte derecha con las tendencias: “Increase in societal source”) o si hay un retraso del benchmark sobre las otras fuentes (parte izquierda “Societal signals to consider”).

**[SOCIETAL SIGNALS TO CONSIDER] - EXPLANATION**

This analytical lens should complement your annual or bi-annual materiality review.

It is based on the data of your analysis as of 06-2020, and provides an additional interpretation of the results by ignoring the source weights applied.

Each column represents a comparison of one or more leading vs. lagging source scores as indicated in the column headers. If the leading source has a higher issues with a 'Total' value of 3 or 4 present stronger signals for consideration; 2 is weaker; 1 and 0 are not presenting.

**[INCREASE IN SOCIETAL SOURCES (6 MONTHS)] - EXPLANATION**

This analytical lens should complement your annual or bi-annual

It assesses the change in the importance of your issues over 6 m

If an issue's importance has grown during that period and accor issues with a 'Total' value of 3 present stronger emergence; 2 is

Issue name	SOCIETAL SIGNALS TO CONSIDER										INCREASE IN SOCIETAL SOURCES (6 MONTHS)			
	External score > Internal score	Newsflow > Benchmark(s)	Voluntary regulation > Mandatory regulation		Social > All other sources	Total	Social	Newsflow	Voluntary regulation	Total	Social	Newsflow	Voluntary regulation	Total
			FALSE	TRUE										
Customer responsibility	TRUE	FALSE	FALSE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1
Employee rights	TRUE	FALSE	FALSE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1	TRUE	TRUE	TRUE	3
Employment practices	FALSE	TRUE	FALSE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	2
Energy efficiency	FALSE	TRUE	TRUE	TRUE	FALSE	2	FALSE	FALSE	FALSE	0	FALSE	FALSE	FALSE	0
Fair and inclusive workplace	TRUE	FALSE	FALSE	TRUE	FALSE	2	FALSE	FALSE	FALSE	0	FALSE	FALSE	FALSE	0
Governance	TRUE	FALSE	FALSE	TRUE	FALSE	2	FALSE	FALSE	FALSE	0	FALSE	FALSE	FALSE	0
Human rights	TRUE	TRUE	TRUE	TRUE	TRUE	4	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1
Inclusion & accessibility	FALSE	TRUE	TRUE	TRUE	FALSE	2	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1
Innovation & digitalization	TRUE	TRUE	FALSE	FALSE	FALSE	2	TRUE	FALSE	FALSE	2	TRUE	TRUE	TRUE	2
Local community support	TRUE	TRUE	FALSE	FALSE	FALSE	2	TRUE	FALSE	FALSE	2	TRUE	TRUE	TRUE	3
Natural capital	TRUE	FALSE	TRUE	TRUE	FALSE	2	FALSE	FALSE	FALSE	0	FALSE	FALSE	FALSE	0
Pollution	TRUE	FALSE	FALSE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1
Responsible investment	FALSE	FALSE	TRUE	TRUE	FALSE	1	FALSE	FALSE	FALSE	1	FALSE	TRUE	TRUE	1
Product Stewardship	FALSE	FALSE	FALSE	FALSE	FALSE	0	TRUE	FALSE	FALSE	2	TRUE	FALSE	FALSE	2
Solvency and economic value	FALSE	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	FALSE	0	FALSE	TRUE	TRUE	2
Transparency	TRUE	FALSE	TRUE	TRUE	FALSE	2	TRUE	FALSE	FALSE	2	TRUE	FALSE	FALSE	2
Waste management	TRUE	FALSE	FALSE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	2
Water	FALSE	FALSE	TRUE	TRUE	FALSE	1	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	2
Wellbeing, health & safety	TRUE	TRUE	FALSE	FALSE	FALSE	2	TRUE	FALSE	FALSE	2	TRUE	TRUE	TRUE	3
Climate change	TRUE	TRUE	TRUE	TRUE	FALSE	3	FALSE	FALSE	FALSE	2	FALSE	TRUE	TRUE	2
Customer privacy and information security	FALSE	FALSE	FALSE	FALSE	FALSE	0	TRUE	FALSE	FALSE	0	TRUE	TRUE	TRUE	3
Employee development	FALSE	FALSE	FALSE	FALSE	FALSE	0	TRUE	FALSE	FALSE	1	TRUE	FALSE	FALSE	1
ESG governance structure	FALSE	TRUE	TRUE	TRUE	FALSE	2	TRUE	FALSE	FALSE	2	TRUE	FALSE	TRUE	2
Skilled workforce	FALSE	FALSE	FALSE	FALSE	FALSE	0	TRUE	FALSE	FALSE	0	TRUE	FALSE	FALSE	1

Ilustración 77 - Ejemplo de tabla en la hoja principal del archivo Excel exportado desde Datamaran, conteniendo la detección de señales y futuras tendencias. (Fuente: Datamaran)



## 6.6. Conclusión

El proyecto de mejora del módulo de Datamaran se planificó a principios de 2020, con el objetivo de ser desplegado el 22 de abril durante un *Webinar* de lanzamiento con el objetivo de presentar el módulo a los clientes. Los recursos principales para este proyecto fueron sin duda el tiempo del personal de diferentes niveles de la organización para elaborar los requerimientos y cumplirlos. Este trabajo trato de cubrir sobre todo la parte de elaboración de señales para identificar y presentar las temáticas indicando una futura tendencia, o un posible incremento en importancia. La pandemia de Covid-19 que ocurrió justo durante el periodo de desarrollo del módulo obviamente tuvo un impacto sobre la planificación, pero la facilidad de trabajar de manera remota y la buena gestión de la crisis por la empresa permitió cumplir con la fecha límite sin mayores problemas.

Al ser una empresa de SaaS<sup>30</sup>, Datamaran no es responsable de gestionar sus servidores, sino que sus proveedores (AWS y MongoDB entre otros) se encargan de gestionar los recursos informáticos necesarios al funcionamiento de la plataforma. Por tanto, la gestión de los problemas se ve simplificada, cómo se puede observar en el AMFE del proyecto. Gracias a varias funcionalidades de Amazon, es fácil programar alarmas que avisan al equipo cuando algo está pasando, y la resolución suele ser rápida.

Es difícil hacer estimaciones del retorno sobre inversión del trabajo, pero con aproximaciones razonables se puede estimar que los futuros nuevos clientes suscribirán al abono anual de Datamaran al menos por una parte gracias al nuevo módulo de *monitoring*. Por tanto, el trabajo será rápidamente rentable cuando se cobrarán los abonos anuales (1 año máximo).

---

<sup>30</sup> Software-as-a-Service





## 7. Conclusiones

### 7.1. Conclusión de la memoria

Para concluir este trabajo, Datamaran provee a sus clientes (responsables de sostenibilidad para multinacionales por la mayoría) una herramienta que permite analizar en profundidad los tópicos actuales más importante. Gracias a una tecnología basada en inteligencia artificial, su algoritmo puede leer e interpretar millones de páginas de texto para extraer una medida de la importancia de ciertos conceptos, y evaluar de forma cuantitativa la prioridad de aquellos a través de varias fuentes (competidores, leyes y medios sociales). Pero si este análisis completo se puede efectuar en un momento dado, los clientes están más interesados en saber de qué manera las temáticas interesantes están evolucionando y sobre todo cuáles serán las más importantes mañana.

En un mundo en cambio constante y dónde la humanidad se queda sorprendida e ineficiente frente a una pandemia como la de 2020, es poco razonable esperar que un algoritmo sencillo sea capaz de predecir el futuro. No obstante, los clientes necesitan una herramienta que les dé más información sobre lo que está pasando de forma que reduzcan su incertidumbre.

Por tanto, la implementación de un modelo sencillo de detección de señales sirve para llamar la atención de cualquier persona que quiere vigilar y gestionar sus riesgos no financieros. Basado en observaciones sencillas sobre la forma general de los datos recolectados por Datamaran, el nuevo de *continuous monitoring* implantado por este trabajo permitirá que los clientes puedan focalizar más su atención sobre las temáticas que realmente representan un riesgo futuro.

En la espera de tener un seguimiento más amplio y personalizado de cada empresa, para poder elaborar un modelo más preciso, usando redes neuronales, por ejemplo, la solución propuesta por el presente estudio permite generar resultados significativamente mejores que una detección aleatoria.

Por tanto, el valor generado por aquella mejora probablemente convencerá a un grupo de clientes potenciales que estaban esperando a que la solución de Datamaran gane en valor añadido. En el escenario más probable para el año 2020, el número de clientes que decidirán suscribir a Datamaran o renovar su suscripción de los años anteriores será más que suficiente para cubrir los gastos necesarios para el trabajo presente.

### 7.2. Lecciones aprendidas

Desde el punto de vista académico, este trabajo me ha permitido aplicar parte de la teoría aprendida en clases de dirección de proyectos, dirección de operaciones, optimización y sistemas de información a un caso concreto y práctico. Por ello, he mejorado mi conocimiento de conceptos como los planes de contingencias, el método Scrum y el método AHP. El lenguaje Python también fue parte de mi carrera en Francia y desarrollar este trabajo me permitió mejorar mis conocimientos del lenguaje y sus bibliotecas.

Profesionalmente, durante mis prácticas en Datamaran y el desarrollo de este trabajo, tuve que aprender a comunicar mis resultados al resto del equipo de la mejor forma posible. Es una competencia entre las más importantes cuando se investigan soluciones un poco complejas, porque muchas veces el cliente o el colega no tiene tiempo, ni conocimientos técnicos suficiente para una explicación completa de la solución. Tuve que aprender a presentar los resultados de la forma más sintética posible y seleccionar lo más relevante. También tuve la oportunidad de trabajar con personas de diferentes áreas en Datamaran y aprender más sobre los diferentes aspectos del marketing y del entrenamiento de los usuarios, por ejemplo.

Aprendí que es importante estudiar durante mucho tiempo los datos hasta entenderlos antes de proponer una solución, porque en el área de la ciencia de datos muchas soluciones funcionan en situaciones restringidas, pero se exportan mal al mundo real. Los conocimientos y la experiencia adquirida con este trabajo aportarán a mi resumen algo que será útil cuando trabajar con inteligencia artificial será una necesidad.

Desde el punto de vista personal, quería realizar un trabajo que tenía sentido para mi, y creo que los valores y la misión de Datamaran me corresponden perfectamente. En efecto, el trabajo de Datamaran pone la tecnología al servicio de las consideraciones éticas y medio ambientales, en un universo donde se tiene que revolucionar el modelo económico para integrar que los recursos disponibles existen en cantidad limitada. Siento que este trabajo me ha permitido poner mi pasión (analizar y manipular los datos) al servicio de una mejor gestión de las temáticas éticas importantes hoy en día.

### 7.3. Futuras líneas de trabajo

Acerca del módulo en sí, se tendrá que preparar una solución que cumpla con el requisito R13 (este ha sido abandonado para el despliegue de abril). Para ello se pretende construir primero un archivo Excel que recopila todos los elementos (informes corporativos, artículos de prensa, texto de ley etc..) que provocaron el incremento de importancia de una temática. Para ello se necesita replicar la lógica de cálculo de Datamaran al revés para determinar, a partir de una temática dada cuáles son las causas de incremento de un momento a otro. También será necesario clasificar las causas según su relevancia para proveer la información más completa posible al usuario, para que pueda contestar fácilmente a la pregunta: ¿por qué mi temática ha subido rangos?

Ahora mismo se está preparando la versión 5 de la plataforma, que tendrá un diseño y una organización completamente diferente. En esta versión se integrará la funcionalidad que se acaba de mencionar, más diferentes filtros y posibilidad de sortear las temáticas según varios criterios. También se cambiará la ontología (definición de los conceptos que los tópicos encajan) para ponerla al día de 2020, y se cambiará la arquitectura completa del módulo para optimizar el tiempo consumido cuando se llama a la base de datos.

El modelo desarrollado en este trabajo para la predicción de tendencias es bastante simple, lo que ayuda al usuario a entender mejor cómo funciona el *software*. Las soluciones evaluadas no son capaces de realizar una predicción correcta en 90% de los casos, pero así constituyen una base para explorar soluciones más complejas que se comparan a las definidas en el trabajo. Pero ahora que esta base está bien hecha, es necesario explorar otras soluciones. Primero se tendría que replicar el estudio con otros datos para confirmar que las alternativas seleccionadas realmente son eficaces con otros entornos. Segundo, con los datos adicionales, intentar otros métodos de predicciones usando horizontes más largos (como por ejemplo usando medias móviles). En caso de obtener resultados mejores con esas soluciones, se podría construir un conjunto de datos más amplio para entrenar un modelo de neuronas artificiales que pueda aprender a predecir automáticamente las futuras tendencias.

El modelo de predicción de momento solo es binario (la temática sube o no sube) pero, se podría intentar predecir una cantidad de rangos de incremento, por ejemplo, o franjas, para incrementar su granularidad. Para poder llegar a un nivel de predicción correcto en más del 90% de los casos también será necesario hacer un análisis en profundidad de las razones (eventos importantes en el mundo) para las cuales una temática u otra puede haber subido. Tras esta identificación se podrán buscar las señales anteriores al evento (si existen) que podrían anunciar el evento con antelación, o por lo menos incrementar la probabilidad de aquel evento.



# Bibliografía

- Bankia. (13 de Diciembre de 2019). *Bankia estudios*. Obtenido de Bankia estudios: <https://www.bankiaestudios.com/estudios/es/publicaciones/la-inversion-con-criterios-esg-esta-de-moda-pero-no-va-a-pasar-de-moda.html>
- Datamaran Ltd. (8 de June de 2018). *Seven tips for materiality analysis*. Obtenido de Datamaran: <https://www.datamaran.com/blog/seven-tips-perfect-materiality-analysis/>
- Datamaran Ltd. (2020). *Datamaran*. Obtenido de About Us: <https://www.datamaran.com/about-us/>
- Label ISR. (s.f.). *Label ISR*. Obtenido de lelabelisr.fr: <https://www.lelabelisr.fr/en/how-label-awarded/>
- Molinares, I. (1 de Octubre de 2019). *CNN Español*. Obtenido de <https://cnnespanol.cnn.com>: <https://cnnespanol.cnn.com/video/trump-ventila-su-molestia-por-juicio-politico-falsas-teorias-ione-molinares/>
- Bruque, M. (6 de Marzo de 2020). *El Español*. Obtenido de <https://www.elespanol.com>: [https://www.elespanol.com/ciencia/medio-ambiente/20200306/calentamiento-global-febrero-calido-seco-espana/472703122\\_0.html](https://www.elespanol.com/ciencia/medio-ambiente/20200306/calentamiento-global-febrero-calido-seco-espana/472703122_0.html)
- León, F. M. (27 de Marzo de 2020). *Tiempo*. Obtenido de <https://www.tiempo.com>: <https://www.tiempo.com/ram/la-calidad-del-aire-en-europa-mejora-por-el-coronavirus.html>
- Saaty, T. L. (2008). *Decision making with the analytic hierarchy process*. Pittsburgh: University of Pittsburgh.
- Gopinath, G. (14 de abril de 2020). *Blog IMF*. Obtenido de <https://blogs.imf.org>: <https://blogs.imf.org/2020/04/14/the-great-lockdown-worst-economic-downturn-since-the-great-depression/>
- (2020). Obtenido de Datamaran: <https://youtu.be/NB3JrqzA9nc>
- Lean Solutions. (s.f.). *AMEF Análisis de Modo y Efecto de Falla*. Obtenido de Lean Solutions: <https://leansolutions.co/conceptos-lean/lean-manufacturing/amef-analisis-de-modo-y-efecto-de-falla/>
- Faculty of George Town. (s.f.). Obtenido de Faculty of GeorgeTown: <https://faculty.georgetown.edu/wilson/IR/WD3.html>
- Rodríguez Abellán, C. (10 de Julio de 2019). *La Ley de Zipf, o por qué hay – casi – el doble de Garcías que de González*. Obtenido de Think Big: [https://empresas.blogthinkbig.com/la-ley-de-zipf/#:~:text=Es%20decir%2C%20este%20patr%C3%B3n%20\(o,de%20la%20distribuci%C3%B3n%20de%20Pareto\).](https://empresas.blogthinkbig.com/la-ley-de-zipf/#:~:text=Es%20decir%2C%20este%20patr%C3%B3n%20(o,de%20la%20distribuci%C3%B3n%20de%20Pareto).)
- Towards Data Science. (4 de Febrero de 2020). *Medium*. Obtenido de Top 10 In-Demand programming languages to learn in 2020: <https://towardsdatascience.com/top-10-in-demand-programming-languages-to-learn-in-2020-4462eb7d8d3e>
- Madnight GitHub. (23 de Junio de 2020). *Madnight GitHub*. Obtenido de [https://madnight.github.io/github/#/pull\\_requests/2020/1](https://madnight.github.io/github/#/pull_requests/2020/1)
- Proyectos Ágiles. (s.f.). *Proyectos Ágiles*. Obtenido de <https://proyectosagiles.org/que-es-scrum/>
- Stamatis, D. H. (2003). *Failure Mode and Effect Analysis: FMEA from Theory to Execution*. Milwaukee, Wisconsin: ASQ Quality Press.
- Newman, M. E. (2006). *Power laws, Pareto distribution and Zipf's law*. Ann Arbor: University of Michigan.
- Schwaber, K., & Sutherland, J. (2017). *The Scrum Guide*. Obtenido de Scrum Guides: <https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf>
- Sutherland, J., & Coplie, J. (2019). *A Scrum Book: The Spirit of the Game*. Pragmatic Bookshelf.



# Anejos

## Anejo A – Código Python Parte 1

24/06/2020

Investigation

```
In [23]:  
  
import pandas as pd  
import numpy as np  
from collections import defaultdict  
import json  
import calendar  
from datetime import datetime, date  
from dateutil.relativedelta import relativedelta  
from IPython.display import display as disp  
from dmaws.s3_export import dataframe_to_excel  
import os  
  
COMPANIES = os.listdir("company_data/post_fix/")[1:]  
  
BENCHMARKS = [f'benchmark{i}' for i in range(5)]  
DM_SOURCES = ['newsflow', 'social', 'observeV', 'observeM'] + BENCHMARKS  
  
ECONOMIC = [100,90,70]  
CONSISTENCY = [60,100,20]  
SELECTIVIDAD_OPTIMA = 90#%  
LAMBDA_STAGE = 'uat'  
WEIGHTS = pd.Series(index=['reliability', "selectividad", "poder predictivo", "coherencia", 'economi  
c'], data=[.49, .19, .19, .09, .04])  
COEFF = [[1], [1,1,1,1], [1]]  
THRESH = [[0.12], [0.1,0.03,0.03,0], [0.12]]  
LIMITE_SELECTIVIDAD = 60#% minimum  
  
def get_lambda_cm_input(company):  
    try:  
        file = open(f'company_data/post_fix/{company}/{company}.json')  
    except:  
        print(f'Could not open {company} file for company in lambda')  
        return  
    return json.loads(file.read())  
  
def lambda_cm(company):  
    lambda_url = LAMBDA.get(LAMBDA_STAGE).get('url')  
    body_company = get_lambda_cm_input(company)  
    headers_lambda_prod= {"Content-Type": "application/json", "x-api-key": LAMBDA.get(LAMBDA_STAGE  
)}.get('key') }  
    resp = requests.post(lambda_url,  
        data=json.dumps(body_company), headers = headers_lambda_prod)  
    return resp.json()  
  
def sort_by_date(mat_data):  
    for month_dict in mat_data:  
        year = month_dict.get('date').get('year')  
        month = month_dict.get('date').get('month')+1  
        month_dict['dt_obj'] = datetime_obj = date(year=year,  
            month=month,  
            day=calendar.monthrange(year=year, month=month)[-1])  
    mat_data.sort(key=lambda x:x.get('dt_obj'))  
  
def evaluate(alternative_num, useful_gaps):  
    if alternative_num -1 not in range(len(ALTERNATIVES)):  
        print('Unknown alternative, should be in', list(range(len(alternatives))))  
        return  
    study = defaultdict(lambda: [])  
    for company in useful_gaps:  
        for table_data_frame, increases_df in useful_gaps.get(company):  
            announce_df = ALTERNATIVES[alternative_num-1](table_data_frame)[f'alternativa{altern  
ative_num}']
```

localhost:8888/nbconvert/html/Investigation.ipynb?download=false

2/16

24/06/2020

Investigation

```
        study[company].append((announce_df, increases_df))
    return study

def performance(study, consistency, economic):
    results = pd.DataFrame()
    for comp in study:
        for ind, gap in enumerate(study.get(comp)):
            temp = {indic:val for indic, val in zip(['reliability', "selectividad", "poder predicti
vo"],
                                                    analyse(gap))}
            results = pd.concat([results, pd.DataFrame(index = [f"{comp}{ind}"], data=temp)])
    results['coherencia'] = [consistency for i in range(len(results))]
    results['economic'] = [economic for i in range(len(results))]
    return results

def analyse(gap):
    total_issues = len(gap[0])
    announced = gap[0].value_counts().get(True, 0)
    confirmed = [gap[0][ind] and gap[1][ind] for ind in range(len(gap[0])].count(True)
    disconfirmed = [not gap[0][ind] and not gap[1][ind] for ind in range(len(gap[0])].count(Tru
e)
    selectividad = round(100*(total_issues-announced)/total_issues, 1)
    selectividad_score = round(100-(abs(selectividad - SELECTIVIDAD_OPTIMA)*100/SELECTIVIDAD_OPT
IMA), 1)
    # disp(gap[0], gap[1])
    if announced:
        poder_predictivo = round(100*confirmed/announced, 1)
    else:
        poder_predictivo = 0
    reliability = 0
    if announced!= total_issues:
        reliability = 100*disconfirmed/(total_issues-announced)
    return reliability, selectividad, poder_predictivo

def load_inputs(company_list=COMPANIES, time_window=6):
    lambda_inputs = defaultdict(lambda: [])
    useful_gaps = defaultdict(lambda: [])
    observations = pd.DataFrame()
    for company in company_list:
        lambda_inputs[company] = get_lambda_cm_input(company)
        mat_data = lambda_inputs.get(company).get('mat_data')
        sort_by_date(mat_data)
        for ind, month_dict in enumerate(mat_data):
            if ind+time_window<len(mat_data):
                future_df = pd.DataFrame(mat_data[ind+time_window].get('table_data')).set_index(
'name')
                table_data_frame = pd.DataFrame(month_dict.get('table_data')).set_index('name')
                increases_df = future_df['stakeholder_score']-table_data_frame['stakeholder_scor
e']>0
                useful_gaps[company].append((table_data_frame, increases_df))
        observations = pd.concat([observations, table_data_frame.loc[increases_df, :]])
    return useful_gaps, observations

def alternativo1(table_data_frame):
    dm_sources = set(table_data_frame.columns).intersection(DM_SOURCES)
    dinamic_sources = {'newsflow', 'social'}
    non_dinam_sources = dm_sources.difference(dinamic_sources)
    table_data_frame["alternativo1"] = [min([table_data_frame.loc[issue, dinam_source] for dinam
source in dinamic_sources]) \
    > THRESH[0][0]+max([table_data_frame.loc[issue, dinam_source] for dinam_source in non_d
inam_sources])\
    for issue in table_data_frame.index]
    return table_data_frame

def alternativa2(table_data_frame):
```

localhost:8888/nbconvert/html/Investigation.ipynb?download=false

3/16



24/06/2020

Investigation

```
signals = pd.DataFrame(index=table_data_frame.index)
dm_sources = set(table_data_frame.columns).intersection(DM_SOURCES)
benchmarks = set([f"benchmark{i}" for i in range(5)]).intersection(set(table_data_frame.columns))
signals['news signal'] = [table_data_frame.loc[issue, 'newsflow'] > THRESH[1][0]+max([table_data_frame.loc[issue,source] for source in benchmarks])\
for issue in table_data_frame.index]
signals['voluntary signal'] = [table_data_frame.loc[issue, 'observeV'] > THRESH[1][1]+max([table_data_frame.loc[issue,source] for source in benchmarks])\
for issue in table_data_frame.index]
signals['social signal'] = [table_data_frame.loc[issue, 'social'] > THRESH[1][2]+max([table_data_frame.loc[issue,source] for source in dm_sources.difference({'social'})])\
for issue in table_data_frame.index]
signals['internal external'] = [table_data_frame.loc[issue, 'stakeholder_score'] > THRESH[1][3]+table_data_frame.loc[issue, 'strategy_score']\
for issue in table_data_frame.index]
table_data_frame['alternativa2'] = [signals.loc[issue,:].value_counts().get(True,0)>=3 for issue in signals.index]
# disp(table_data_frame)
return table_data_frame
def alternativa4(table_data_frame):
dm_sources = set(table_data_frame.columns).intersection(DM_SOURCES)
dinamic_sources = {'benchmark', 'social'}
benchmarks = set([f"benchmark{i}" for i in range(5)]).intersection(set(table_data_frame.columns))
non_dinam_sources = dm_sources.difference(dinamic_sources)
table_data_frame["alternativa3"] = [max([table_data_frame.loc[issue,bench] for bench in ['benchmark0']) \
>= 0
# table_data_frame.loc[issue, 'observeV'] > THRESH[2][1]+table_data_frame.loc[issue, 'stakeholder_score']
for issue in table_data_frame.index]
return table_data_frame
def alternativa3(table_data_frame):
dm_sources = set(table_data_frame.columns).intersection(DM_SOURCES)
dinamic_sources = {'benchmark', 'social'}
benchmarks = set([f"benchmark{i}" for i in range(5)]).intersection(set(table_data_frame.columns))
non_dinam_sources = dm_sources.difference(dinamic_sources)
table_data_frame["alternativa3"] = [max([table_data_frame.loc[issue,bench] for bench in ['benchmark0']) \
> THRESH[2][0]+table_data_frame.loc[issue, 'stakeholder_score']\
# table_data_frame.loc[issue, 'observeV'] > THRESH[2][1]+table_data_frame.loc[issue, 'stakeholder_score']
for issue in table_data_frame.index]
return table_data_frame
def presence(df, exclude=[]):
df_numbers = df.select_dtypes('number')
exclude = df.columns.difference(df_numbers.columns).union(exclude)
include = df.columns.difference(exclude)
return pd.concat([(100*df_numbers/df_numbers.sum())[include], df[exclude]], axis=1)
ALTERNATIVES = [alternativa1, alternativa2, alternativa3]
In [227]:
USEFUL_GAPS, OBSERVATIONS = load_inputs()
```

localhost:8888/nbconvert/html/Investigation.ipynb?download=false

4/16

## Anejo B – Código Python Parte 2

24/06/2020

Investigation

```
In [229]:  
for alternative_num in range(len(ALTERNATIVES)):  
    scores = performance(evaluate(alternative_num+1,USEFUL_GAPS),CONSISTENCY[alternative_num],EC  
ONOMIC[alternative_num]).mean()  
    total_score = (scores*WEIGHTS).sum()  
    print(scores,'Total score: ',total_score)  
  
reliability      71.372647  
selectividad     96.843333  
poder predictivo 27.221667  
coherencia       60.000000  
economic         100.000000  
dtype: float64 Total score: 67.94494686922076  
reliability      70.235612  
selectividad     89.561667  
poder predictivo 28.388333  
coherencia       100.000000  
economic         90.000000  
dtype: float64 Total score: 69.42595005776778  
reliability      75.330126  
selectividad     71.988333  
poder predictivo 36.638333  
coherencia       20.000000  
economic         70.000000  
dtype: float64 Total score: 62.15082852973443
```

```
In [22]:  
for alternative_num in range(len(ALTERNATIVES)):  
    scores = performance(evaluate(alternative_num+1,USEFUL_GAPS),CONSISTENCY[alternative_num],EC  
ONOMIC[alternative_num]).mean()  
    total_score = (scores*WEIGHTS).sum()  
    print(scores,'Total score: ',total_score)  
  
fiabilidad       28.795745  
exactitud        6.433333  
selectividad_score(90) 92.970175  
selectividad     92.843860  
scope_reduction  5.615086  
economic         90.000000  
dtype: float64 Total score: 23.101592385218368  
fiabilidad       26.870175  
exactitud        10.263158  
selectividad_score(90) 92.303509  
selectividad     86.740351  
scope_reduction  8.509111  
economic         100.000000  
dtype: float64 Total score: 25.956824561403515  
fiabilidad       32.098246  
exactitud        76.896491  
selectividad_score(90) 34.345614  
selectividad     30.915789  
scope_reduction  23.553647  
economic         60.000000  
dtype: float64 Total score: 65.2063859649123
```

24/06/2020

Investigation

```
In [225]:  
# optimize param  
THRESH = [[0],[0,0,0,0],[0]]  
total_score = [[] for i in range(3)]  
coeffs = [[],[],[]]  
epsilon = 0.01  
MAX_ITE = 11  
step = 0.01  
for alternative_num in range(len(ALTERNATIVES)):  
    for thresh_ind in range(len(THRESH[alternative_num])):  
        i=0  
        scores = performance(evaluate(alternative_num+1,USEFUL_GAPS),CONSISTENCY[alternative_num],  
ECONOMIC[alternative_num]).mean()  
        overall_score=round((scores*WEIGHTS).sum(),1)  
        print('Alternativa',alternative_num+1,scores,overall_score,THRESH[alternative_num][thresh_ind])  
#         penul_score = total_score[alternative_num][-1]  
#         if overall_score >0:  
total_score[alternative_num].append(overall_score)  
coeffs[alternative_num].append(THRESH[alternative_num].copy())  
#         last_score = 0  
while i<=MAX_ITE:  
    i+=1  
    THRESH[alternative_num][thresh_ind]+=step  
    scores = performance(evaluate(alternative_num+1,USEFUL_GAPS),CONSISTENCY[alternative_num],  
ECONOMIC[alternative_num]).mean()  
    overall_score=round((scores*WEIGHTS).sum(),1)  
    print('Alternativa',alternative_num+1,scores,overall_score,THRESH[alternative_num][thresh_ind])  
#         penul_score = 0  
#         if len(total_score[alternative_num])>1:  
#             penul_score = total_score[alternative_num][-2]  
last_score = total_score[alternative_num][-1]  
if overall_score >last_score:  
    total_score[alternative_num].append(overall_score)  
    coeffs[alternative_num].append(THRESH[alternative_num].copy())  
#         penul_score = last_score  
#         last_score = overall_score  
    print(total_score[alternative_num],coeffs[alternative_num])  
    THRESH[alternative_num][thresh_ind]=coeffs[alternative_num][-1][thresh_ind]  
# total_score[alternative_num] = total_score[alternative_num][-1]  
# coeffs[alternative_num] = coeffs[alternative_num][-1]  
print(f"Solution for alternative {alternative_num+1}: score({total_score[alternative_num][-1]})"+  
f" threshs({coeffs[alternative_num][-1]})")  
THRESH[alternative_num][thresh_ind] = coeffs[alternative_num][-1]
```

localhost:8888/nbconvert/html/Investigation.ipynb?download=false

6/16



24/06/2020

Investigation

```

Alternativa 1 reliability          71.009543
selectividad                      92.365000
poder predictivo                  25.556667
coherencia                        60.000000
economic                          100.000000
dtype: float64 66.6 0
Alternativa 1 reliability          71.077725
selectividad                      92.643333
poder predictivo                  25.556667
coherencia                        60.000000
economic                          100.000000
dtype: float64 66.7 [0.01]
[66.6, 66.7] [[0], [0.01]]
Alternativa 1 reliability          71.077725
selectividad                      92.643333
poder predictivo                  25.556667
coherencia                        60.000000
economic                          100.000000
dtype: float64 66.7 [0.02]
Alternativa 1 reliability          71.117687
selectividad                      93.540000
poder predictivo                  25.585000
coherencia                        60.000000
economic                          100.000000
dtype: float64 66.9 [0.03]
[66.6, 66.7, 66.9] [[0], [0.01], [0.03]]
Alternativa 1 reliability          71.299295
selectividad                      94.296667
poder predictivo                  26.833333
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.4 [0.04]
[66.6, 66.7, 66.9, 67.4] [[0], [0.01], [0.03], [0.04]]
Alternativa 1 reliability          71.413939
selectividad                      94.650000
poder predictivo                  26.833333
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.5 [0.05]
[66.6, 66.7, 66.9, 67.4, 67.5] [[0], [0.01], [0.03], [0.04], [0.05]]
Alternativa 1 reliability          71.399670
selectividad                      94.866667
poder predictivo                  26.750000
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.5 [0.060000000000000005]
Alternativa 1 reliability          71.288131
selectividad                      95.051667
poder predictivo                  26.750000
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.5 [0.07]
Alternativa 1 reliability          71.321465
selectividad                      95.116667
poder predictivo                  26.750000
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.5 [0.08]
Alternativa 1 reliability          71.203860
selectividad                      95.458333
poder predictivo                  26.528333
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.5 [0.09]
Alternativa 1 reliability          71.281048
selectividad                      95.865000
poder predictivo                  26.666667
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.6 [0.09999999999999999]
[66.6, 66.7, 66.9, 67.4, 67.5, 67.6] [[0], [0.01], [0.03], [0.04], [0.05], [0.09999999999999999]]
Alternativa 1 reliability          71.391586
selectividad                      96.710000
poder predictivo                  26.666667
coherencia                        60.000000
economic                          100.000000
dtype: float64 67.8 [0.10999999999999999]
[66.6, 66.7, 66.9, 67.4, 67.5, 67.6, 67.8] [[0], [0.01], [0.03], [0.04], [0.05], [0.09999999999999999], [0.10999999999999999]]
Alternativa 1 reliability          71.372647
selectividad                      96.843333
poder predictivo                  27.221667
coherencia                        60.000000
economic                          100.000000

```

localhost:8888/nbconvert/html/Investigation.ipynb?download=false

7/16

## Anejo C – Código Python Parte 3

24/06/2020

Investigation

```
In [35]:  
THRESH
```

```
[[0.16], [0.06, 0.16, 0.16, 0.16], [0.16, 0.16]]
```

```
In [ ]:  
observations= OBSERVATIONS.copy()  
# observations = observations[observations.columns.intersection(DM_SOURCES)]  
observations['benchmark(max)'] = observations[observations.columns.intersection(BENCHMARKS)].max  
(axis=1)  
observations['benchmark(avg)'] = observations[observations.columns.intersection(BENCHMARKS)].mea  
n(axis=1)  
observations['observe sig'] =observations['benchmark0']>observations['stakeholder_score']  
observations['observe sig'].value_counts()[True]/len(observations)
```

```
In [ ]:  
#alternativ: coeff=0.65 score = 24.2
```

```
In [24]:  
SOCIAL_SOURCES = ['observeV', 'social', 'newsflow']  
def load_inputs2(company_list=COMPANIES,time_window=6,trend_months=3):  
    lambda_inputs = defaultdict(lambda: [])  
    useful_gaps = defaultdict(lambda: [])  
    data_dict = defaultdict(lambda:defaultdict(lambda:pd.DataFrame()))  
    for company in company_list:  
        lambda_inputs[company] = get_lambda_cm_input(company)  
        mat_data = lambda_inputs.get(company).get('mat_data')  
        sort_by_date(mat_data)  
        for ind,month_dict in enumerate(mat_data):  
            if ind+time_window+trend_months<len(mat_data):  
                future_df = pd.DataFrame(mat_data[ind+time_window+trend_months].get('table_data'  
)).set_index('name')
```

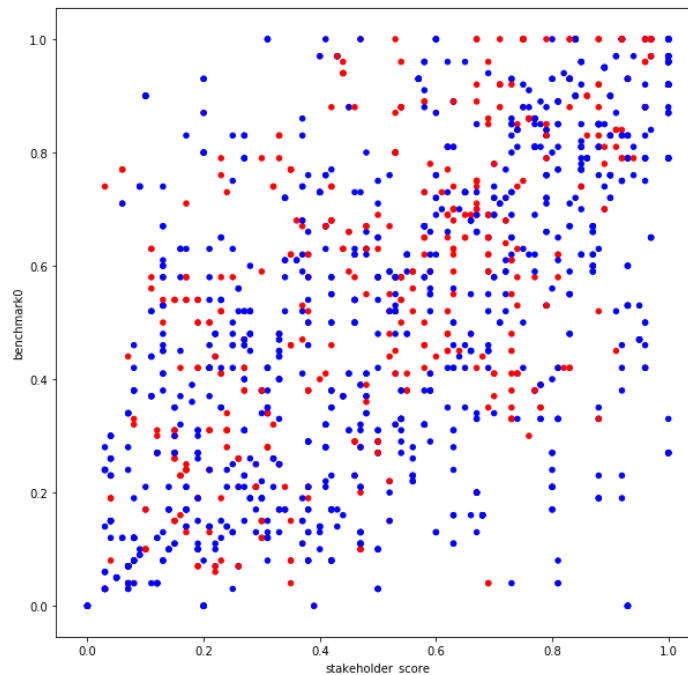
localhost:8888/nbconvert/html/Investigation.ipynb?download=false

15/16

24/06/2020

Investigation

```
In [56]:  
DATA_SET = pd.DataFrame()  
for company in USEFUL_GAPS:  
    for data,label in USEFUL_GAPS.get(company):  
        data['label'] = label  
        data['color_label'] = ['blue' for i in range(len(data))]  
        data.loc[data['label'],'color_label'] = ['red' for i in range(len(data.loc[data['label']  
, 'color_label']))]  
        data['analysis'] = [company] * len(data)  
  
        data_analysis = data.copy()  
        data_analysis.reset_index(inplace=True)  
        DATA_SET = pd.concat([DATA_SET,data_analysis])  
DATA_SET = DATA_SET.fillna("")  
# disp(DATA_SET)  
DATA_SET.plot.scatter(x='stakeholder_score',y='benchmark0',c=DATA_SET['color_label'],figsize=(10  
,10))  
with pd.ExcelWriter("export.xlsx") as writer:  
    dataframe_to_excel(workbook=writer.book,data_df=[DATA_SET],top_text = ['DATA SET'],sheet_nam  
e='data')
```



24/06/2020

Investigation

```
table_data_frame = pd.DataFrame(mat_data[ind+time_window].get('table_data')).set_index('name')
issue_data_frame0 = pd.DataFrame(month_dict.get('issue_data'),index=future_df.index)
issue_data_frame1 = pd.DataFrame(mat_data[ind+time_window].get('issue_data'),index=future_df.index)
increases_df = future_df['stakeholder_score']-table_data_frame['stakeholder_score']>0
useful_gaps[company].append((issue_data_frame0,issue_data_frame1,increases_df))
observations = pd.concat([observations,table_data_frame.loc[increases_df,:]])
return useful_gaps,observations

def load_inputs3(company_list=COMPANIES,time_window=6,trend_confirm=3):
    lambda_inputs = defaultdict(lambda: [])
    useful_gaps = defaultdict(lambda:[])
    data_dict = defaultdict(lambda:pd.DataFrame())
    with pd.ExcelWriter('increasing_trends.xlsx',engine='xlsxwriter') as writer:
        for source in SOCIAL_SOURCES:
            for company in company_list:
                lambda_inputs[company] = get_lambda_cm_input(company)
                mat_data = lambda_inputs.get(company).get('mat_data')
                sort_by_date(mat_data)
                for ind,month_dict in enumerate(mat_data):
                    if ind+time_window+trend_confirm+1<len(mat_data):
                        index = pd.DataFrame(mat_data[ind]['table_data'])['name']
                        month_ind = -1
                        company_data = pd.DataFrame()
                        for month_int in list(range(ind,ind+time_window+1))+[ind+time_window+trend_confirm-1]:
                            monthly_presence = presence(pd.DataFrame(data=mat_data[month_int].get('issue_data'),
                                                                    index=index))
                            # month_str=str(mat_data[month_int].get('date').get('year'))+str(mat_data[month_int].get('date').get('month')+1).zfill(2)
                            # disp(monthly_presence)
                            new_data = monthly_presence[source]
                            month_ind+=1
                            new_data.name = str(month_ind)
                            if month_int == ind+time_window+trend_confirm-1:
                                month_ind+=2
                                new_data = monthly_presence[source]>company_data[company_data.columns[-1]]
                                new_data.name = "label"
                                company_data = pd.concat([company_data,new_data],axis=1)
                            company_data['analysis'] = [company for i in range(len(company_data))]
                            data_dict[source]=pd.concat([data_dict[source],company_data.reset_index()])
                        dataframe_to_excel(workbook=writer.book,data_df=[data_dict[source].fillna("")],sheet_name=source,top_text=[source])
    return data_dict

In [25]:
INCREASING_GAPS = load_inputs3()
```

localhost:8888/nbconvert/html/Investigation.ipynb?download=false

16/16



# Anejo D – Ontología: definición de los tópicos

<p><b>ECONOMIC</b></p> <p><b>Digitalization</b> References to the optimization of business activities, processes, and models through the use of digital technology. E.g. artificial intelligence, virtual currency, fintech</p> <p><b>Impact investing products</b> References to financial products based on or with a goal of promoting environmental, social and governance (ESG) themes. E.g. green bonds, sustainability oriented private equity and insurance</p> <p><b>Long-term value</b> References to creation of value for shareholders and society in the long term. E.g. long term growth, direct and indirect economic impact, shared value</p> <p><b>Market access</b> References to sell and/or to sell off resources requiring access of goods and services to national markets. E.g. free trade agreements, embargoes, trade sanctions</p> <p><b>Responsible investment and financing</b> References to measures emphasizing the inclusion and striving to generate both financial and sustainable value. E.g. green investing, impact investing, ethical funds</p> <p><b>Responsible marketing</b> References to marketing for communication, promotion, and selling of products and/or services. E.g. cause-related marketing, marketing to vulnerable groups</p> <p><b>Responsible pricing</b> References to the application of pricing policies for products and services. E.g. affordability products, services, and solutions, fair pricing, transparent prices</p> <p><b>Responsible procurement</b> References to the acquisition of goods or services from an external source in a responsible manner. E.g. green sourcing policies, procurement ethics, purchasing policies</p> <p><b>Solvency and financial management</b> References to the financial state of a company and the ability to meet long term and short term financial obligations. E.g. capital budget, charge performance, solvency, liquidity</p>	<p><b>ENVIRONMENTAL</b></p> <p><b>Forests</b> References to forest resources. E.g. timber certification, deforestation, forest restoration</p> <p><b>Greenhouse gases</b> References to gases trapping heat in the atmosphere causing the rise in global temperature (greenhouse effect). E.g. CO2, carbon management systems, carbon capture systems</p> <p><b>Harmful substances</b> References to substances that are used in the operations and/or present in the products posing a substantial or potential threat to public health or the environment. E.g. toxins, carcinogenic substances, dangerous chemicals</p> <p><b>Hazardous waste</b> References to waste posing a substantial or potential threat to public health or the environment. E.g. toxic waste, cleaning waste, radioactive waste</p> <p><b>Land</b> References to use and development of land resources in both urban and rural settings. E.g. land conservation, urbanization, land restoration</p> <p><b>Materials management</b> References to the use and development of materials and resources in the operations, activities, products, or materials development and use of commodities, such as timber, paper</p> <p><b>Noise pollution</b> References to unwanted, excessive, and/or harmful level of noise that impacts a community or larger ecosystem. E.g. sleep disturbance, hearing and general chronic hearing damage</p> <p><b>Non-hazardous waste</b> References to waste that is not considered hazardous. E.g. industrial waste, water management, water treatment</p> <p><b>Product stewardship</b> References to a product's environmental impacts during its lifecycle. E.g. life cycle assessment, product durability, product take back</p> <p><b>Recycling, waste management &amp; reduction</b> References to the management of waste management, reduction and/or conversion into reusable material. E.g. recycling programs and initiatives</p> <p><b>Renewables &amp; alternatives</b> References to energy from natural processes and/or non-renewable resources that are replenished on a human timescale. E.g. alternative energy sources, photovoltaic, biomass</p> <p><b>Spills</b> References to substances (discharge of substances) on an ecosystem. E.g. oil spill, chemical release, spill prevention and recovery</p> <p><b>Sustainable building management</b> References to environmental management of buildings and facilities. E.g. building efficiency, sustainable design, smart buildings</p> <p><b>Water pollution</b> References to water bodies affected in terms of quality due to use or contamination. E.g. groundwater contamination, sewage, water quality, freshwater contamination</p> <p><b>Water recycling, efficiencies &amp; remediation</b> References to recycling, reuse and/or remediation of water. E.g. water reclamation, water recycling, water remediation</p> <p><b>Water resources</b> References to water availability and distribution. E.g. water scarcity, surface and ground water, desalination</p>	<p><b>SOCIAL</b></p> <p><b>Community support &amp; development</b> References to well-being, support, and development of community, geographical area, or defined group of people. E.g. community education, support to local business, community safety</p> <p><b>Conflict minerals</b> References to minerals sourced from countries or regions in conditions of armed conflict and human rights abuses. E.g. minerals associated with war zones, conflict-free certification schemes</p> <p><b>Consumer rights</b> References to consumer's right to safe, informed choices, and give input to the producers of the goods and services they purchase. E.g. consumer information, actions, and protection</p> <p><b>Customer privacy &amp; information security</b> References to customer and other confidential information protection, integrity and availability. E.g. cyber terrorism, network security, identity theft</p> <p><b>Customer satisfaction</b> References to meeting or surpassing customer expectations. E.g. customer engagement, feedback, and loyalty</p> <p><b>Digital inclusion</b> References to the ability of individuals and/or groups to access and use information and communication technologies (ICT). E.g. digital literacy, digital divide, and ICT coverage to remote areas</p> <p><b>Disaster management, recovery &amp; relief</b> References to social and political processes and/or initiatives. E.g. disaster preparedness and response, humanitarian aid, disaster management</p> <p><b>Disease awareness and education</b> References to ensuring effective medical treatments through promotion of health literacy and understanding of diseases. E.g. health literacy, patient education, disease prevention</p> <p><b>Energy inclusion</b> References to the ability of individuals and/or groups to access energy resources. E.g. affordable energy, fuel poverty, energy justice</p> <p><b>Fair competition</b> References to just and equitable competition based on factors of price, quality and services. E.g. monopolies and cartels, price-fixing, anti-trust</p> <p><b>Financial inclusion</b> References to accessibility of appropriate financial services, products, and resources. E.g. microfinance institutions, fair design of financial products</p> <p><b>Financial literacy</b> References to the ability of individuals and/or groups to use financial products and services. E.g. financial education, finance awareness, financial capabilities</p> <p><b>Forced labour</b> References to work or services which people are forced to do against their will under the threat of some form of punishment. E.g. slavery, human exploitation, unauthorised work</p> <p><b>Geopolitical risk</b> References to political or regional instability due to political positions, tensions, international relations. E.g. terrorism, international relations, civil rights, discrimination</p> <p><b>Human rights</b> References to the rights of all people, based on fundamental rights ensuring basic conditions for individuals to dignity. E.g. fundamental freedoms, civil rights, discrimination</p> <p><b>Human trafficking</b> References to the movement of people against their will for the purposes of forced labor or commercial sexual exploitation. E.g. smuggling of people, forced migration</p> <p><b>Indigenous populations</b> References to communities that are geographically distinct traditional habitats or ancient territories, and who identify themselves as part of a distinct cultural group. E.g. free prior and informed consent, indigenous rights, indigenous territories</p> <p><b>Innovation, research &amp; development</b> References to the technical development of new products and services. E.g. new patents, robotics, research partnerships</p> <p><b>Land rights &amp; displacement</b> References to the rights of a community or lands impacted by the business presence, including the settlement of compensation due to business activities. E.g. land expropriation, voluntary and involuntary displacement</p> <p><b>Local economy</b> References to production, distribution and consumption activities in the local economic system. E.g. local sourcing and procurement, local employment, local infrastructure</p> <p><b>Net neutrality</b> References to an internet service provider's treatment of data and traffic. E.g. neutral treatment of internet data, favoring websites or networks, digital freedom</p> <p><b>Nutrition</b> References to food, nutrients, dietary needs. E.g. healthy food, substances in food, nutritional information</p> <p><b>Philanthropy</b> References to voluntary provision of resources, time or other inputs to promote welfare of others. E.g. charity, donations, non-profit support</p> <p><b>Product &amp; service safety</b> References to health and safety risks related to the use of products and/or services. E.g. consumer health, safe use of products, digital products</p> <p><b>Product labelling</b> References to information displayed on products, including its origin, contents and suggested usage. E.g. transparency of composition/ingredients, labeling requirements, packaging information</p> <p><b>Reputation</b> References to positive or negative perception of an entity by stakeholders based on objective assessment of its conduct and ability to deliver on its commitments. E.g. corporate image, brand value, business development</p> <p><b>Sexual exploitation</b> References to sexual or non-sexual exploitation of different power or trust to sexual purposes. E.g. sex slavery, genital mutilation, child prostitution</p> <p><b>Social inclusion</b> References to the ability of individuals and groups to take part in society. E.g. non-discrimination, social mobility, vulnerable and diverse groups</p> <p><b>Stakeholder inclusion</b> References to the inclusion of groups who may affect or be affected by the entity's activities and decisions. E.g. stakeholder engagement, involvement and dialogue</p>
<p><b>EMPLOYEE</b></p> <p><b>Eco-efficient employee transportation</b> References to environmental friendly commuting of employees. E.g. green bicycles, fleet, alternative carbon transportation</p> <p><b>Employee attraction &amp; retention</b> References to attract and/or retention of current employees with skills and skills. E.g. talent scouting and recruiting, human capital management</p> <p><b>Employee compensation &amp; benefits</b> References to financial and/or non-financial compensation offered to employees by their employer. E.g. remuneration policies and programs, pension plans, health care support</p> <p><b>Employee development</b> References to activities and programs to advance skills of employees. E.g. career advancement programs, scholarships, coaching and mentoring programs</p> <p><b>Employee engagement</b> References to employee commitment, motivation, and participation in the organization life. E.g. feedback from employees, employee surveys, employee surveys</p> <p><b>Employee satisfaction</b> References to the degree to which employees are satisfied with their employment and employment conditions. E.g. employee morale and happiness, diversity of workplace</p> <p><b>Employee transportation</b> References to employee commuting and travel. E.g. workplace mobility and transportation</p> <p><b>Employee volunteering</b> References to the voluntary contribution of time and skills to contribute to social and community development during work time. E.g. volunteer programs, volunteer management</p> <p><b>Employee wellbeing</b> References to mental, emotional, psychological and physical conditions of employees in their workplace. E.g. work-life balance, harassment, absenteeism</p> <p><b>Fair remuneration</b> References to fair and equitable pay for the same work. E.g. equitable remuneration, gender pay gap</p> <p><b>Labor rights</b> References to legal rights requiring employers to follow labor rules. E.g. working conditions, parental leave, maternity protection</p> <p><b>Occupational health &amp; safety</b> References to health, safety and well-being of employees and others in the workplace. E.g. incident and safety, accident investigation, ergonomics</p> <p><b>Skilled workforce</b> References to workforce with high level of expertise and know-how, characterized by high education or expertise level. E.g. skilled labor, high-potential employees</p> <p><b>Unionization</b> References to employees' collective action. E.g. formation of unions, union organizing, strikes</p> <p><b>Workforce changes</b> References to addition or removal of people employed by an entity. Includes changes due to restructuring, reorganization, turnover, etc. E.g. restructuring, turnover, etc.</p> <p><b>Workforce diversity &amp; inclusion</b> References to incorporating a diverse workforce and equal opportunity for employees of diverse backgrounds. E.g. equal opportunities and treatments, inclusion workplace, harassment</p>	<p><b>GOVERNANCE</b></p> <p><b>Anti-corruption &amp; bribery</b> References to activities that are prohibited, such as the offering, giving, receiving, or holding of something of value for the purpose of influencing an action. E.g. corporate governance, bribery, money laundering</p> <p><b>Board composition</b> References to the composition of highest executive body of an entity. E.g. board diversity, gender equity, and independence</p> <p><b>Business ethics</b> References to the management and moral code of conduct in the strategic and operational management of business. E.g. corporate ethics, fiduciary duty, code of conduct</p> <p><b>Corporate culture</b> References to shared values, beliefs and behaviors that characterize an entity and define its identity. E.g. entity's vision and mission statements, core principles, and corporate identity</p> <p><b>ESG Governance structure</b> References to governance bodies and the structure of managing and overseeing an organization's ESG dimensions. E.g. ESG CSR/ Sustainability departments, committees, directors</p> <p><b>Executive compensation</b> References to financial and/or non-financial compensation to highest level of senior management. E.g. executive bonus, golden parachute, stock benefits</p> <p><b>Grievance Mechanisms</b> References to systems to receive and respond to reports of non-compliance with the company's policies, procedures, or standards. E.g. whistleblowing, anonymous reporting, dispute resolution</p> <p><b>Investor relations</b> References to the relationship and their shareholders, current and potential investors, financial analysts. E.g. shareholders engagement, financial public relations</p> <p><b>IP infringement</b> References to the violation of intellectual property or moral rights protecting the products of human intelligence and creation, such as patents and trademarks. E.g. violation of patents, trademarks, copyrights</p> <p><b>Non-financial accounting</b> References to the systematic recording, recording and analysis of business transactions beyond financial lines. E.g. environmental indicators, trust ratings, Global Reporting Initiative</p> <p><b>Public policy practices</b> References to influencing the production, enactment, execution and interpretation of legislation. E.g. governmental relations, lobbying activities, political contributions</p> <p><b>Responsible tax practices</b> References to lawful tax payment. E.g. responsible tax, tax transparency, tax haven</p> <p><b>Shareholder activism</b> References to shareholders' power and rights to influence company policies. E.g. active ownership, writing general resolutions, shareholder resolutions</p> <p><b>Supply chain engagement</b> References to involvement of third parties with suppliers. E.g. collaboration throughout the supply chain, supplier participation</p> <p><b>Supply chain management</b> References to supply chain activities, including system and downstream flow of resources and information. E.g. supply chain quality, suppliers, sales</p>	<p><b>ENVIRONMENTAL</b></p> <p><b>Air emissions</b> References to emissions and associated greenhouse gas that impact air quality and/or atmospheric conditions. E.g. air pollutants, smog, particulate</p> <p><b>Alternative fuels</b> References to fuel derived from non-fossil resources with a lower environmental impact. E.g. alternative fuels, biofuels, hydrogen or natural gas</p> <p><b>Animal welfare</b> References to animal health and well-being and values of animals. E.g. animal testing, animal rights, animal care</p> <p><b>Biodiversity</b> References to biodiversity, i.e. genetic or species variety of plants and animals present within an area, biome or planet. E.g. protected species and land, wildlife, conservation initiatives</p> <p><b>Climate change</b> References to human-induced changes in the distribution of weather patterns. E.g. climate events, sea level rise, climate change</p> <p><b>Eco-efficient transportation</b> References to human-induced changes in the distribution of weather patterns. E.g. car sharing and pooling, walking to work, cycling to work</p> <p><b>Emission trading</b> References to government-mandated and market-based strategies for mitigating climate change. E.g. green trading, clean trading, and low carbon transportation</p> <p><b>Energy reduction &amp; efficiencies</b> References to government-mandated and market-based strategies for mitigating climate change. E.g. green trading, clean trading, and low carbon transportation</p> <p><b>Energy use</b> References to consumption of renewable &amp; non-renewable forms of energy by activities, operations, products. E.g. power usage, fuel efficiency, lighting automation</p> <p><b>Fisheries</b> References to fishery resources. E.g. responsible fishing, aquaculture, marine biodiversity</p>



Datamaran Benchmark & Observe topics, as of January 7th, 2019  
© 2014-2019 Datamaran. All rights reserved.  
www.datamaran.com

# Presupuesto

## Contenido

- Consideraciones previas
- Justificación de precios por capítulos
- Presupuesto de ejecución material
- Presupuesto por contrata



## Consideraciones previas

Este proyecto es un proyecto de investigación e implantación de herramienta enteramente informática, no se ha usado material otro que ordenadores, y material básico de oficina. El recurso principal necesario será equivalente a los salarios de las personas que contribuyeron al proyecto.

Este proyecto, aunque se implementa realmente en la empresa Datamaran, es un Trabajo de Fin de Máster y por tanto se le ha dedicado como mínimo el tiempo de trabajo equivalente a 12 créditos ECTS o 300h de trabajo.

### Servicios de Amazon

Cómo se ha detallado el capítulo sobre los requerimientos de la nueva herramienta, se van a usar funcionalidades de AWS<sup>31</sup>: las lambdas y el espacio de almacenamiento S3. El uso de aquellos tiene un coste variable en función de los que se usa:

- La lambda es una función que se invoca y tiene acceso a los recursos del cloud (o servidor) durante su tiempo de ejecución y luego se apaga. Su coste mensual depende de la memoria viva que necesita para ejecutar sus tareas, del tiempo de ejecución más un coste fijo por invocación. El precio de computación para cada GB.s de ejecución será de  $1,48 \cdot 10^{-5} \text{€}$ , y el precio por invocación será de  $1,78 \cdot 10^{-7} \text{€}$ . Además, para una suscripción a los servicios AWS, cada mes, los primeros 400 000 GB.s y las primeras 1M invocaciones serán gratis entre todas las lambdas usadas.
- Los *buckets S3* son repositorios para almacenar datos en el servidor, seguros y accesibles. El coste es de 0,0205€/GB almacenado, los primeros 5GB son gratis.

Durante los 6 primeros meses de 2020, cuando se hicieron las fases de test y desarrollo de la herramienta y entonces se usó de manera significativa el módulo (más que después de su despliegue al público) los datos picos fueron los siguientes:

- Lambda: 31 100 invocaciones, usando un total de 116 500 GB.s en un mes. El coste mensual de uso con esta misma frecuencia de uso sería de unos 2€/mes. Incluyendo todas las otras lambdas de Datamaran (incluso las que no se usan para este módulo) no se alcanza el límite de los servicios gratis de AWS.
- S3 *Buckets*: máximo 1,7 GB ocupados solamente por los archivos del módulo de *monitoring*. Esto representa unos 0,03€/mes que también se despreciara.

Es probable que en la vida futura de la empresa el uso del módulo pase a ser relevante a la hora de calcular costes. Con 60 clientes que usan la plataforma aproximadamente 2 veces al año, incluso si este número dobla antes del fin del año los costes de computación y almacenamiento se quedarán por debajo del límite de la oferta de Amazon.

### Base de datos (MongoDB)

Acerca del almacenamiento de datos en las bases de datos gestionadas por MongoDB, no se ha cambiado la estrategia, no se necesitó más espacio de lo que se usaba antes para el módulo anterior, por tanto, se considerará un coste nulo.

### Sueldos del personal

Como los costes de material (a parte de lo indirecto) se desprecian todos, el coste del proyecto será enteramente el tiempo del personal que contribuyó a la elaboración del proyecto,

---

<sup>31</sup> Amazon Web Service

tomando como base los sueldos siguientes:

CTO	50.00 €
Customer success	35.00 €
Desarrollador	30.00 €
Director de producto	35.00 €
Ingeniero de datos	25.00€
Ingeniero de datos senior	35.00 €

#### **Costes indirectos**

Se considerará un 1% adicional en costes indirectos, es decir costes necesarios a la realización del proyecto, pero relacionados indirecta o difícilmente evaluables.

#### **Gastos generales**

En concepto de gastos generales se considerará un 10% adicional del presupuesto de ejecución material por los siguientes costes:

- Material de oficina
- Energía
- Amortización del ordenador usado
- Conexión internet

#### **Beneficio industrial**

Se considerará un beneficio industrial del 6% del valor del presupuesto de ejecución material, como beneficios en caso de que una empresa privada realice este proyecto para que genere beneficios.

## Justificación de precios por capítulo

Capítulo/Subcapítulo	Código	Ud.	Descripción	Precio unitario	Importe
<b>1. Introducción</b>					
1.1 Objeto del trabajo	P01.01	1 h	Ingeniero de datos	25.00 €	25.00 €
		1%	<i>Costes Indirectos</i>	25.00 €	0.25 €
			<b>Total</b>		<b>25.25 €</b>
1.2 Metodología	P01.02	5 h	Ingeniero de datos	25.00 €	125.00 €
		1%	<i>Costes Indirectos</i>	125.00 €	1.25 €
			<b>Total</b>		<b>126.25 €</b>
1.3 Antecedentes teóricos	P01.03	10 h	Ingeniero de datos	25.00 €	250.00 €
		1%	<i>Costes Indirectos</i>	250.00 €	2.50 €
			<b>Total</b>		<b>252.50 €</b>
1.4 Estructura del documento	P01.04	2 h	Ingeniero de datos	25.00 €	50.00 €
		1%	<i>Costes Indirectos</i>	50.00 €	0.50 €
			<b>Total</b>		<b>50.50 €</b>
<b>2. Descripción del entorno</b>					
2.2 La empresa Datamaran Evalue SLU	P02.01	10 h	Ingeniero de datos	25.00 €	250.00 €
		1%	<i>Costes Indirectos</i>	250.00 €	2.50 €
			<b>Total</b>		<b>252.50 €</b>
2.3 Los clientes y proveedores de la empresa	P02.02	5 h	Ingeniero de datos	25.00 €	125.00 €
		1%	<i>Costes Indirectos</i>	125.00 €	1.25 €
			<b>Total</b>		<b>126.25 €</b>
2.4 Productos y servicios de la empresa	P02.03	20 h	Ingeniero de datos	25.00 €	500.00 €
		1%	<i>Costes Indirectos</i>	500.00 €	5.00 €
			<b>Total</b>		<b>505.00 €</b>

Capítulo/Subcapítulo	Código	Ud.	Descripción	Precio unitario	Importe
<b>3. Análisis de la situación actual</b>					
3.2 Desarrollo de un análisis de materialidad	P03.01	30 h	Ingeniero de datos	25.00 €	750.00 €
		1%	<i>Costes Indirectos</i>	750.00 €	7.50 €
			<b>Total</b>		<b>757.50 €</b>
3.3 Continuous monitoring	P03.02	10 h	Ingeniero de datos	25.00 €	250.00 €
		1%	<i>Costes Indirectos</i>	250.00 €	2.50 €
			<b>Total</b>		<b>252.50 €</b>
<b>4. Definición de las necesidades</b>					
4.2 Requerimientos obligatorios	P04.01	20 h	Ingeniero de datos	25.00 €	500.00 €
		5 h	Ingeniero de datos senior	35.00 €	175.00 €
		5 h	Desarrollador	30.00 €	150.00 €
		3 h	Director de producto	35.00 €	105.00 €
		3 h	CTO	50.00 €	150.00 €
		3 h	Customer success	35.00 €	105.00 €
		1%	<i>Costes Indirectos</i>	105.00 €	1.05 €
			<b>Total</b>		<b>1,186.05 €</b>
4.3 Requerimientos libres	P04.02	30 h	Ingeniero de datos	25.00 €	750.00 €
		4 h	Ingeniero de datos senior	35.00 €	140.00 €
		2 h	Customer success	35.00 €	70.00 €
		2 h	Director de producto	35.00 €	70.00 €
		2 h	CTO	50.00 €	100.00 €
		1%	<i>Costes Indirectos</i>	100.00 €	1.00 €
			<b>Total</b>		<b>1,131.00 €</b>



Capítulo/Subcapítulo	Código	Ud.	Descripción	Precio unitario	Importe
<b>5. Definición de los indicadores</b>					
5.2 Metodología	P05.01	5 h	Ingeniero de datos	25.00 €	125.00 €
		1%	<i>Costes Indirectos</i>	125.00 €	1.25 €
<b>Total</b>					<b>126.25 €</b>
5.3 Señales sociales	P05.02	100 h	Ingeniero de datos	25.00 €	2,500.00 €
			Ingeniero de datos		
		10 h	senior	35.00 €	350.00 €
		2 h	CTO	50.00 €	100.00 €
		1%	<i>Costes Indirectos</i>	100.00 €	1.00 €
<b>Total</b>					<b>2,951.00 €</b>
5.4 Tendencias	P05.03	120 h	Ingeniero de datos	25.00 €	3,000.00 €
			Ingeniero de datos		
		5 h	senior	35.00 €	175.00 €
		1 h	CTO	50.00 €	50.00 €
		1%	<i>Costes Indirectos</i>	50.00 €	0.50 €
<b>Total</b>					<b>3,225.50 €</b>
<b>6. Implantación de la herramienta</b>					
6.2 Plan de implantación	P06.01	20 h	Ingeniero de datos	25.00 €	500.00 €
		1%	<i>Costes Indirectos</i>	500.00 €	5.00 €
<b>Total</b>					<b>505.00 €</b>
6.3 Plan de contingencias	P06.02	25	Ingeniero de datos	25.00 €	625.00 €
		1%	<i>Costes Indirectos</i>	625.00 €	6.25 €
<b>Total</b>					<b>631.25 €</b>
6.4 Valoración económica del proyecto	P06.03	10	Ingeniero de datos	25.00 €	250.00 €
		1%	<i>Costes Indirectos</i>	250.00 €	2.50 €
<b>Total</b>					<b>252.50 €</b>



## Presupuesto de ejecución material

Capítulo/Subcapítulo	Importe
1. Introducción	454.50 €
2. Descripción del entorno	883.75 €
3. Análisis de la situación actual	1,010.00 €
4. Definición de las necesidades	2,317.05 €
5. Definición de los indicadores	6,302.75 €
6. Implantación de la herramienta	1,388.75 €
<b>Presupuesto ejecución material</b>	<b>12,356.80 €</b>

Importe del presupuesto de ejecución material: DOCE MIL TRESCIENTOS CINCUENTA Y SEIS EUROS CON OCHENTA CÉNTIMOS

Ingeniero Industrial  
Vincent Rizzo



## Presupuesto por contrata

Capítulo/Subcapítulo	Importe	
1. Introducción		454.50 €
2. Descripción del entorno		883.75 €
3. Análisis de la situación actual		1,010.00 €
4. Definición de las necesidades		2,317.05 €
5. Definición de los indicadores		6,302.75 €
6. Implantación de la herramienta		1,388.75 €
<b>Presupuesto ejecución material</b>		<b>12,356.80 €</b>
Gastos generales	10%	1,235.68 €
Beneficio industrial	6%	741.41 €
<b>Presupuesto por contrata</b>		<b>14,333.89 €</b>
IVA	21%	3,010.12 €
<b>Presupuesto por contrata con IVA</b>		<b>17,344.00 €</b>

Importe del presupuesto por contrata (IVA incluido): DIECISIETE MIL TRESCIENTOS CUARENTA Y CUATRO EUROS

Ingeniero Industrial  
Vincent Rizzo