



Escuela Técnica Superior de Ingeniería Informática

Ingeniería de Software

TRABAJO FIN DE GRADO

SM-Meter: Una ventana a las redes sociales.

Autor/es:

Pérez Sosa, Ezequiel

Tutor/es:

Ortega Rodríguez, Francisco Javier

Primera convocatoria

Curso <2021/2022>

Índice

Tabla de Contenido	5
Tabla de Ilustraciones.....	6
Agradecimientos	8
Resumen.....	9
Capítulo 1.	10
Alcance	10
1.1. Enunciado del alcance	10
1.2. Objetivos	10
1.3. Límites	11
1.4. Proceso de verificación del alcance.	11
1.5. Documentación de Requisitos.....	11
Interesados.....	13
Tecnologías.....	14
Metodologías	15
1.1. Metodología a utilizar	15
1.2. Adaptación de Scrum	15
Planificación	16
1.1. Definición	16
1.2. Planificación Temporal	16
1.3. Estimación tareas	17
1.4. Planificación de costos	19
1.4.1. Costes directos	19
1.4.2. Costes indirectos	20
1.4.3. Resumen.....	20
Capítulo 2.	22
Estudio previo	22
1.1. Historia del procesamiento del lenguaje natural.....	22
1.2. Introducción al análisis del sentimiento.....	22

2. Evolución de las tecnologías	23
3. Aporte de este proyecto	25
3.1. Limitaciones y contexto.....	26
Datasets y datos de entrenamiento.....	27
1.1. Qué es y cuáles son los datos objetivos de este proyecto.....	27
1.2. Preprocesamiento de los datasets	28
Entrenamiento de modelos.....	30
1. Elección de modelos pre-entrenados y ajuste fino	30
2. Funcionamiento de Bert.....	30
Modelos entrenados	32
1.1. Subdivisión de tareas	32
1.2. Primeras pruebas	32
1.3. Entrenamiento de los modelos definitivos	33
1.4. Evaluación de los modelos resultantes	37
Capítulo 3.	42
Motivación del desarrollo	42
Tecnologías.....	43
Mockups	44
1. Importancia del diseño.....	44
2. Herramienta utilizada.....	44
3. Ilustraciones y explicación del diseño	45
Desarrollo	49
1. Proceso de desarrollo.....	49
2. Flujo de la aplicación	51
3. Capturas finales.....	52
Conclusiones	55
1. Conclusiones.....	55
Capítulo 4.	56
Como interpretar este capítulo.....	56
Datasets.....	57
Elecciones EE. UU. 2020	58
1. Introducción	58
2. Por qué es interesante estudiar estas elecciones	58
3. Estudio elecciones US 2020.....	59
4. Conclusiones.....	65
Invasión Rusa.....	66

1. Introducción	66
2. Matices del estudio	66
3. Estudio invasión Rusa.....	67
4. Conclusiones.....	75
Capítulo 5.	76
Consideraciones finales.....	76
1. Tiempo empleado en la realización del proyecto.	76
2. Conclusiones generales.....	77
3. Trabajo futuro.	78
ANEXO I – Lista de Interesados	79
ANEXO II – Glosario de Términos	81
ANEXO III – Bibliografía y Referencias.....	87

Tabla de Contenido

Tabla 1- Requisitos de negocio.....	11
Tabla 2 - Requisitos de los interesados	12
Tabla 3 - Interesados (Resumen).....	13
Tabla 4 - Estimación de tareas	19
Tabla 5 - Resumen del salario de perfiles informáticos	19
Tabla 6 - Costes directos	20
Tabla 7 - Costes indirectos	20
Tabla 8 - Resumen de gastos.....	20
Tabla 9 - Especificaciones hardware del sistema de desarrollo.....	26
Tabla 10 - Ejemplo de dataset en español	27
Tabla 11 - Distribución de datos en los modelos entrenados.....	27
Tabla 12 - Listado de datasets usados.....	28
Tabla 13 - Ejemplo de procesamiento de texto	29
Tabla 14 - Hiper parámetros de los modelos	33
Tabla 15 - Resultados modelo twitter de sentimiento.....	37
Tabla 16 - Resultados modelo reddit de sentimiento	38
Tabla 17 - Resultados modelo español de sentimiento	39
Tabla 18 - Resultados modelo twitter de agresividad.....	39
Tabla 19 - Resultados modelo reddit de agresividad	40
Tabla 20 - Resultados modelo español de agresividad	41
Tabla 21 - Suscripción gratuita MockPlus	44
Tabla 22 - Exportación de datos en CSV.....	54
Tabla 23 - Datos utilizados para el análisis.....	57
Tabla 24 - Palabras más frecuentes en las elecciones	60
Tabla 25 - Palabras agrupadas por movimiento social	68
Tabla 26 - Palabras seleccionadas para la gráfica de eventos	69
Tabla 27 - Estimación de tareas	77
Tabla 28 - Lista de interesado	80
Tabla 29 - Glosario de términos	86

Tabla de Ilustraciones

Ilustración 1 - Resumen global cronograma	16
Ilustración 2 - Cronograma: Sprint 1	16
Ilustración 3 - Cronograma: Sprint 2	16
Ilustración 4 - Cronograma: Sprint 3	17
Ilustración 5 - Cronograma: Sprint 4	17
Ilustración 6 - Cronograma: Sprint 5	17
Ilustración 7 - Evolución de las búsquedas de "Sentiment analysis" a lo largo del tiempo.....	23
Ilustración 8 - Composición neuronal del tipo RNN	24
Ilustración 9 - Composición neuronal del tipo RNN	24
Ilustración 10 - Arquitectura de las redes RNN, LSTM y GRU	24
Ilustración 11 –Funcinamiento MLM	30
Ilustración 12 - Procesamiento datasets.....	34
Ilustración 13 - Procesamiento de etiquetas	34
Ilustración 14 - Proceso de tokenización	35
Ilustración 15 - Algoritmo de backpropagation	36
Ilustración 16 - Red neuronal tipo entrenada.....	36
Ilustración 17 - MockPlus: Vista de proyecto.....	44
Ilustración 18 - MockPlus: Vista de elementos	45
Ilustración 19 - SM-Meter Home.....	45
Ilustración 20 - SM-Meter: Resultados de un análisis de CSV.....	46
Ilustración 21 - SM-Meter Análisis en tiempo real	46
Ilustración 22 - SM-Meter Análisis personalizado.....	47
Ilustración 23 - SM-Meter opciones: General.....	47
Ilustración 24 - SM-Meter opciones: API	48
Ilustración 25 - SM-Meter opciones: About.....	48
Ilustración 26 - SM-Meter: Métodos en el archivo main.py	49
Ilustración 27 - SM-Meter: Métodos en el archivo apiSupport.py	50
Ilustración 28 - SM-Meter: Métodos en el archivo neural.py	50
Ilustración 29 - SM-Meter: Diagrama de secuencia	51
Ilustración 30 - APP: SM-Meter Home	52
Ilustración 31 - APP: SM-Meter Análisis en tiempo real	52
Ilustración 32 - APP: SM-Meter Análisis personalizado	53
Ilustración 33 - APP: SM-Meter opciones: About	54
Ilustración 34 - Nube de palabras de las elecciones de EE. UU en Twitter.....	59
Ilustración 35 - Nube de palabras de las menciones a Biden de EE. UU.....	60
Ilustración 36 - Nube de palabras de las menciones a Trump de EE. UU	60
Ilustración 37 - Matriz de correlación de las elecciones de EE. UU.	61
Ilustración 38 - Nº. Tweets de Trump y Biden	61
Ilustración 39 - Ejemplo de comentario hacia uno de los dos candidatos.....	62
Ilustración 40 - Rango de positividad y agresividad de los comentarios recibidos por los candidatos.....	62
Ilustración 41 - Nº. Tweets con geolocalización activada por candidato.....	62

Ilustración 42 - Mapa de resultados tras el análisis de sentimiento.....	63
Ilustración 43 - Mapa de resultados tras el análisis de agresividad.....	63
Ilustración 44 - Mapa de resultados tras el análisis completo.....	64
Ilustración 45 - Resultado final de las elecciones de EE. UU. de 2020.....	64
Ilustración 46- Nube de palabras de los datos de la invasión rusa	67
Ilustración 47 - Distribución de los comentarios por país.....	68
Ilustración 48 - Evolución de los movimientos sociales con el tiempo	68
Ilustración 49 - Matriz de correlación de apoyo a Ucrania	69
Ilustración 50 - Matriz de correlación contra Putin	69
Ilustración 51 - Matriz de correlación a favor de para la guerra.....	69
Ilustración 52 - Evolución de las palabras claves en el tiempo	70
Ilustración 53 - Evolución de las palabras de eventos en el tiempo	70
Ilustración 54 - Evolución del grado de positividad y odio de la palabra Kharkiv.....	71
Ilustración 55 - Evolución del grado de positividad y odio de la palabra Mariúpol	72
Ilustración 56 - Evolución del grado de positividad y odio de la palabra kiev	73
Ilustración 57- Evolución del grado de positividad y odio de la palabra bucha.....	74
Ilustración 58- Evolución del grado de positividad y odio de la palabra nato	75

Agradecimientos

A mis padres, los cuales se han dejado la piel para que yo pueda redactar estas líneas y me han apoyado en cada paso que he dado siempre, por su amor incondicional y por mirar más en sus propios hijos que en sí mismos, demostrándome cómo quiero ser en la vida. A mi hermano por haberme aconsejado en cada decisión que he tomado y animarme en cada tropiezo que he dado para no quedarme abajo, sino salir más fuerte, ser yo.

A mi pareja por haberme ayudado en cada decisión que he tomado y acompañarme durante todo el camino desde que empecé esta etapa académica, demostrando que puedo con más de lo que he creí.

A mi tutor, por guiarme y acompañarme en el camino de este proyecto, por su implicación desinteresada y ayuda en cada paso que he dado en este trabajo.

A mis amigos que me ayudan cada día a ser mejor persona y que me enseñan que el amor no es solo cosa de familia, o bien, que la familia no es solo de sangre.

A mis revisoras por leer un tema que pese a que es complicado han hecho todo el esfuerzo en entenderlo y darme consejos para que esto salga lo mejor posible.

Y por último, a los que me habría gustado que me acompañasen siempre, a mis abuelos que son la verdadera razón por la que hoy estoy aquí disfrutando este fin del camino.

*No. Try not. Do... or do not. There is no try.
Star wars - Yoda*

Resumen

SM - Meter: Una ventana a las redes sociales es un proyecto que utiliza los datos que obtenemos de redes sociales y el Deep Learning para procesar las opiniones de los usuarios para sacar conclusiones sobre estas y medir qué está ocurriendo en ese momento, o bien, respecto a un tema de interés. El objetivo principal es poder medir el sentimiento de agresividad y positividad (positividad se llamará a veces, en adelante, simplemente sentimiento) respecto un comentario de Twitter o Reddit. Se seguirá la metodología Scrum y las tecnologías de Transformers y programación en Python. Al finalizar se concluye, en base a las investigaciones realizadas, si realmente es efectiva esta manera de análisis.

Alcance

1.1. Enunciado del alcance

Con este proyecto se pretende construir una aplicación sobre análisis y procesamiento del lenguaje natural (PLN). El objetivo que se persigue es desarrollar una aplicación que, apoyándose en el campo de PLN, clasifique las emociones y agresividad de un texto escrito tanto en español como en inglés.

Se persigue que esta aplicación pueda:

- Analizar un texto y devolver cómo se percibe en distintas redes sociales.
- Analizar las emociones de una red social en tiempo real.

Basándonos en la aplicación que se ha descrito anteriormente, se analizarán hechos históricos recientes con el fin de comparar cómo ha afectado a las redes y la forma en la que los propios usuarios se han expresado en términos de agresividad y emociones.

Se analizarán las elecciones de EE. UU. De 2020 y la invasión rusa a territorio ucraniano en 2022.

1.2. Objetivos

Los objetivos asociados a la planificación del proyecto serán los siguientes:

- **OBJ-01:** Realización de una aplicación de escritorio
 - **OBJ-01.1:** El sistema deberá permitir al usuario introducir un texto para su posterior análisis.
 - **OBJ-01.2:** El sistema deberá permitir al usuario introducir un fichero CSV para su posterior análisis.
 - **OBJ-01.3** El sistema deberá permitir visualizar al usuario el estado de la red social Twitter y Reddit en tiempo real.
 - **OBJ-01.4:** El sistema deberá permitir al usuario introducir un criterio de búsqueda sobre un producto o persona para su posterior análisis en Twitter.
- **OBJ-02:** Realización del algoritmo de PLN
 - **OBJ-02.1:** Recolección de datos de Twitter y Reddit.
 - **OBJ-02.2:** Procesado de datos recolectados.
 - **OBJ-02.3:** Análisis de los datos recolectados
 - **OBJ-02.4:** Construcción del modelo para el análisis de sentimiento.
 - **OBJ-02.5:** Construcción del modelo para el análisis de agresividad
 - **OBJ-02.6:** Testeo del modelo
 - **OBJ-02.7:** Optimización del modelo
- **OBJ-03:** Análisis de casos recientes

- **OBJ-03.1:** Recolección y análisis de datos en el marco temporal de las elecciones de EE. UU. en 2020.
- **OBJ-03.2:** Recolección y análisis de datos en el marco temporal de la invasión rusa a territorio ucraniano en 2022.

1.3. Límites

Los datos a los que podemos acceder son enteramente públicos y sólo accedemos a lo que el usuario decide compartir para todo el público. El sistema se podría afinar mejor e incluso mejorarlo añadiendo funcionalidades como detección de *cyberbullying* si se tuviesen los datos privados tales como mensajes privados entre usuarios.

La API de Twitter que usaremos es la pública y gratis. Esto quiere decir que podremos obtener 100 *twits* por búsqueda y el número de peticiones por minuto es inferior al servicio de pago. Además solo podemos obtener tweets a partir de una fecha determinada.

1.4. Proceso de verificación del alcance.

Cada vez que se finaliza un sprint, este es revisado por el product owner. En caso de que sea aceptado se da por finalizado, de lo contrario se realizan las modificaciones pertinentes.

1.5. Documentación de Requisitos

Los requisitos de listan a continuación en forma de tabla

Código	Título	Descripción	Prioridad
RN-01	Elaboración algoritmo PLN	Elaboración de un algoritmo que use técnicas PLN que cumpla con los objetivos marcados en el alcance del proyecto	ALTA
RN-02	Elaboración una aplicación de escritorio	Elaboración de una aplicación de escritorio que de soporte al algoritmo PLN.	ALTA
RN-03	Elaboración de informes de análisis	Elaboración de informes de análisis de hechos históricos recientes al hacer uso del algoritmo PLN.	ALTA
RN-04	Elaboración de la memoria del TFG	Elaboración de la memoria que describe el proyecto en cuestión.	MEDIA
RN-05	Presentación del proyecto	Presentación del proyecto realizado.	BAJA

Tabla 1- Requisitos de negocio

Código	Título	Descripción	Prioridad
RI-01	Análisis de sentimientos y agresividad	La aplicación debe permitir al usuario que este pueda analizar un texto usando el algoritmo presentado.	ALTA
RI-02	Análisis de un texto simple	La aplicación debe permitir al usuario analizar un texto simple, eligiendo el tipo de análisis.	ALTA

Código	Título	Descripción	Prioridad
RI-03	Análisis de varios textos en formato CSV	La aplicación debe permitir al usuario que este pueda analizar varios textos al introducir un csv en la aplicación.	MEDIA
RI-04	Información de la tasa de acierto	Información sobre la tasa de acierto del algoritmo con los datos de entrenamiento.	BAJA
RI-05	Análisis de las redes sociales en tiempo real	Posibilidad de analizar las redes sociales de Twitter y Reddit en tiempo real.	MEDIA
RI-06	Informes sobre el marco temporal de las elecciones de EE. UU. en 2020.	Análisis e informe de los datos recogidos sobre las elecciones de EE. UU. de 2020 después de ser procesados por el algoritmo diseñado en el proyecto.	MEDIA
RI-07	Informes sobre el marco temporal de la invasión rusa a territorio ucraniano en 2022.	Análisis e informe de los datos recogidos sobre la crisis y tensiones diplomáticas entre Rusia, Ucrania y la ONU después de ser procesados por el algoritmo diseñado en el proyecto.	MEDIA

Tabla 2 - Requisitos de los interesados

Interesados

En la siguiente tabla se indican los interesados clave en el proyecto, con su rol y forma de contacto.

Código	Nombre	Rol	Comunicación	Correo corporativo
IN-01	Francisco Javier Ortega Rodríguez	Product Owner, Mentor	Correo y reuniones (Tutorías)	javierortega@us.es
IN-02	Ezequiel Pérez Sosa	Participante del equipo de desarrollo	Correo y reuniones	ezepersos@alum.us.es

Tabla 3 - Interesados (Resumen)

Tecnologías

Para este proyecto utilizaremos las siguientes tecnologías para aplicar PLN:

- **TensorFlow:** Para trabajar con redes neuronales y entrenarlas. En concreto trabajaremos con **BERT** para la interpretación del lenguaje que se apoya en redes neuronales. Utilizamos BERT por su bidireccionalidad a la hora de analizar una frase, ya que podemos tomar en cuenta el contexto y trabajar la polisemia de las palabras.
- **HuggingFace:** Para cargar modelos generales pre-entrenados en BERT. De esta manera afinaremos el modelo según nuestro propio contexto.

Metodologías

1.1. Metodología a utilizar

La metodología de trabajo que se llevará a cabo en este TFG estará basada en Scrum.

Esta metodología se aplicará para organizar el proyecto para conseguir sus objetivos y así separar el trabajo en Sprints según el incremento objetivo de estos.

1.2. Adaptación de Scrum

Para facilitar el uso de Scrum con las fechas y la disponibilidad de los integrantes se ha adaptado scrum modificando algunos aspectos de su aplicación.

- Los roles que se utilizarán serán los siguientes:
 - Product Owner: El tutor de este TFG.
 - Project Manager: El alumno que realiza este proyecto.
 - Equipo de desarrollo: El alumno que realiza este proyecto será el único integrante de este equipo.

Las reuniones se han modificado reduciendo su frecuencia y buscando la disponibilidad de los integrantes, intentando en la medida de lo posible efectuarlas una vez por semana o bien una vez por cada tarea compleja finalizada o próxima a finalizar.

Estas reuniones se realizarán presencialmente o bien desde enseñanza virtual utilizando la herramienta *Black Board Collaborate*.

Planificación

1.1. Definición

Entendemos por planificación cómo la estimación temporal y por ende el coste asociado al proyecto que se presenta.

1.2. Planificación Temporal

La planificación se ha realizado usando Microsoft Project. Se han considerado 5 Sprint para la realización del proyecto que se explican a continuación.

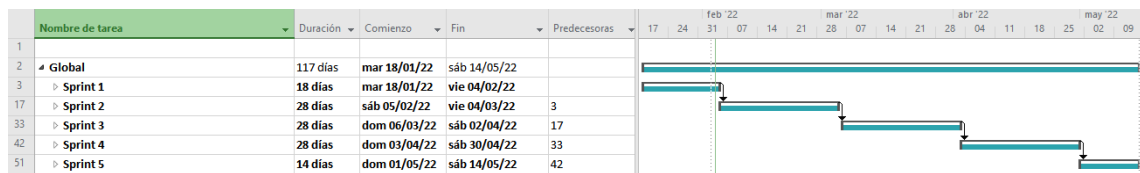


Ilustración 1 - Resumen global cronograma

Como resumen tenemos los 5 Sprint repartidos en el tiempo con un total de 117 días con fecha de comienzo del primer sprint el día 18 de enero de 2022 y el último que con fecha de finalización en el 14 de mayo de 2022.

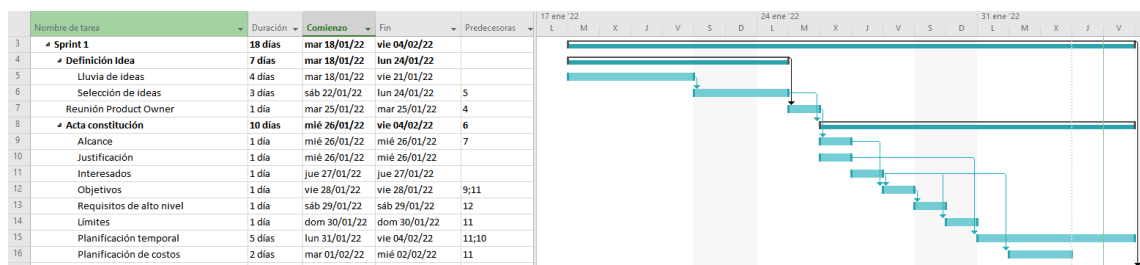


Ilustración 2 - Cronograma: Sprint 1

El primer sprint de 18 días está enfocado a definir el proyecto y realizar el acta de constitución junto con otra documentación importante para el inicio de este.

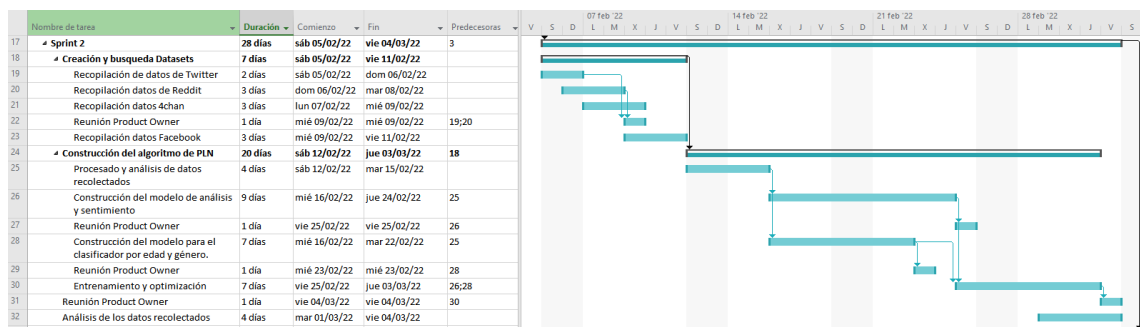


Ilustración 3 - Cronograma: Sprint 2

El segundo sprint de 28 días está enfocado a desarrollar el algoritmo de PLN que será el núcleo del proyecto. Con este sprint se trata de recabar los datos necesarios para el desarrollo y entrenamiento del algoritmo. Finalmente se analizan los resultados obtenidos.

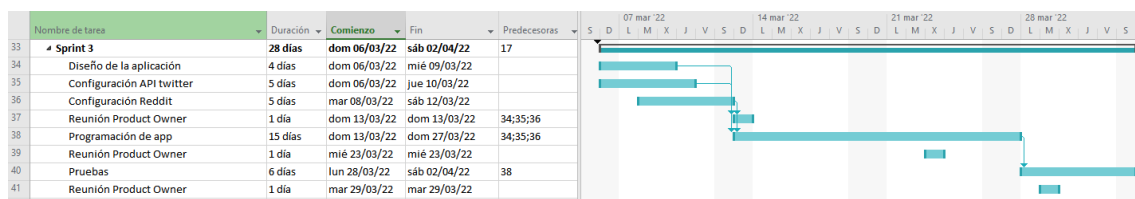


Ilustración 4 - Cronograma: Sprint 3

El tercer sprint de 28 días está enfocado en desarrollar la aplicación destinada al usuario para operar con el algoritmo desarrollado en el anterior sprint. Además, se realizarán más pruebas de la aplicación cómo del algoritmo de forma funcional.

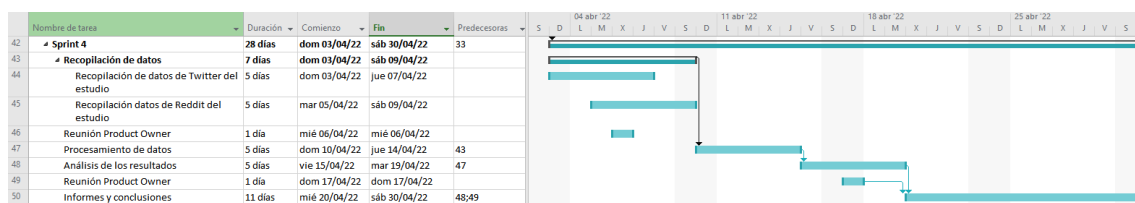


Ilustración 5 - Cronograma: Sprint 4

El cuarto Sprint de 28 días trata de realizar análisis e informes sobre el tercer objetivo de este proyecto, el cual trata de analizar cómo reaccionaron los usuarios en hechos históricos recientes.

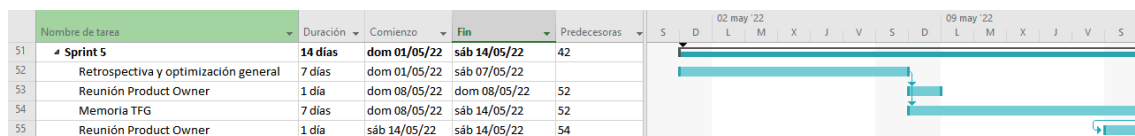


Ilustración 6 - Cronograma: Sprint 5

Finalmente, este último sprint de 14 días se enfoca en optimizar y repasar todo el proyecto completo y finalizar la memoria de este.

1.3. Estimación tareas

	Título tarea	Encargado	D. Optimista	D. Más probable	D. Pesimista
Sprint 1	Lluvia de ideas	J. Proyecto Analista	180	360	480
	Selección de ideas	J. Proyecto Analista	30	120	240
	Reunión Product Owner	J. Proyecto	30	60	120
	Alcance	J. Proyecto Analista	40	120	200

Título tarea		Encargado	D. Optimista	D. Más probable	D. Pesimista
	Justificación	J. Proyecto	30	60	80
	Interesados	J. Proyecto Analista	30	60	80
	Objetivos	J. Proyecto Analista	25	120	160
	Requisitos de alto nivel	Analista	60	120	180
	Límites	J. Proyecto Analista	30	60	80
	Planificación temporal	J. Proyecto	240	480	600
	Planificación de costos	J. Proyecto	60	120	200
Total Sprint			755	1680	2420
Sprint 2	Recopilación de datos de Twitter	Desarrollador	120	300	400
	Recopilación datos de Reddit	Desarrollador	120	300	400
	Reunión Product Owner	J. Proyecto	150	240	300
	Procesado y análisis de datos recolectados	Desarrollador	240	360	500
	Construcción del modelo de análisis y sentimiento	Desarrollador	480	840	920
	Entrenamiento y optimización	Desarrollador	300	540	700
	Análisis de los datos recolectados	Desarrollador	240	360	480
Total Sprint			1650	3060	4150
Sprint 3	Diseño de la aplicación	D. Gráfico	300	360	400
	Configuración API Twitter	Desarrollador	260	420	480
	Configuración Reddit	Desarrollador	300	420	480
	Programación de app	Desarrollador	900	1200	1500
	Reunión Product Owner	J. Proyecto	120	180	200
	Pruebas	Desarrollador	360	540	620
Total Sprint			2240	3120	3680
Sprint 4	Recopilación de datos de Twitter del estudio	Desarrollador	300	480	520
	Recopilación datos de Reddit del estudio	Desarrollador	300	480	520
	Procesamiento de datos	Desarrollador	630	960	980
	Análisis de los resultados	Desarrollador	340	540	600
	Informes y conclusiones	J. Proyecto Desarrollador	500	960	1000
	Reunión Product Owner	J. Proyecto	100	120	180
Total Sprint			2170	3540	3800

Título tarea		Encargado	D. Optimista	D. Más probable	D. Pesimista
Sprint 5	Retrospectiva y optimización general	J. Proyecto Desarrollador Analista	480	1200	1400
	Memoria TFG	J. Proyecto	720	1440	1600
	Reunión Product Owner	J. Proyecto	80	120	180
Total Sprint			1280	2760	3180
Total estimación en horas (Redondeado)			135h	236h	287h

Tabla 4 - Estimación de tareas

1.4. Planificación de costos

Para la realización de la estimación de costes dividiremos estos en dos tipos:

- Costes directos: Asociados al producto sin hacer ningún tipo de reparto
- Costes indirectos: Asociados al proceso de producción en general sin ser asignados a un solo producto o sin criterio de asignación.

1.4.1. Costes directos

Nos referimos como costes directos a la mano de obra que clasificaremos según los roles de cada uno de los miembros.

Basándonos en el documento titulado “PERFILES PROFESIONALES ÁMBITO INFORMÁTICO” que se encuentra dentro de la plataforma de gestión de TFGs de la escuela superior de ingeniería informática (ETSII) calculamos los costes directos del proyecto.

Rol	Salario bruto anual	Salario bruto mensual (12 meses)	Salario bruto por hora
Jefe de proyecto	75,187.20€	6.265.60€	39.16€
Analista	55,142.40€	4,595.20€	28.72€
Desarrollador	46,272.00€	3,856.00€	24.10€
Diseñador	52,934.40€	4,411.20€	27.57€

Tabla 5 - Resumen del salario de perfiles informáticos

Para calcular el coste de cada Sprint consideraremos las horas de las tareas que se han especificado previamente en la tabla de las tareas junto con su encargado. Las horas han sido redondeadas a favor del trabajador.

Sprint	Concepto	Horas totales – Coste/Hora	Coste
Primero	Salario jefe de proyecto	26 horas - 39.16€	1,018.16€
	Salario Analista	16 horas - 28.72€	459.52€
Total Sprint			1,477.68€
Segundo	Salario jefe de Proyecto	4 horas – 39.16€	156.64€
	Salario Desarrollador	69 horas - 24.10€	1,662.90€

Sprint	Concepto	Horas totales – Coste/Hora	Coste
Total Sprint			1,819.54€
Tercero	Salario jefe de Proyecto	3 horas – 39.16€	117.48€
	Salario Desarrollador	43 horas – 24.10€	1036.30€
	Salario Diseñador	6 horas – 27.57€	165.42€
Total Sprint			1,319.20€
Cuarto	Salario jefe de Proyecto	18 horas – 39.16€	704.88€
	Salario Desarrollador	57 horas – 24.10€	1,373.7€
Total Sprint			2,078.58€
Quinto	Salario jefe de Proyecto	46 horas – 39.16€	1,801.36€
	Salario Desarrollador	20 horas - 24.10€	482.00€
	Salario Analista	20 horas - 28.72€	574.40€
Total Sprint			2,857.76€
Total costes directos			9,552.76€

Tabla 6 - Costes directos

1.4.2. Costes indirectos

Hacemos referencia a los costes indirectos desglosados en tres tipos:

- Servicios: Aquellos contratados temporalmente mientras dura la realización del proyecto, tales como el servicio de internet, electricidad, agua, etc.
- Material fungible: Material consumible usado durante el desarrollo del proyecto, tales como cualquier tipo de material de oficina.
- Material no fungible: Material que al ser usado no supone una consumición de este, hacemos referencia por ejemplo a los equipos informáticos.

Concepto	Coste mensual
Alquiler local	600€
Contratación eléctrica	65€*
Contratación agua	20.88€
Servicio Internet (600Mb Simétricos)	29.95€
Coste total tras 4 meses	2863.32€

Tabla 7 - Costes indirectos

* Precio medio de la tarifa PVPC en el 2021

1.4.3. Resumen

TIPO	COSTE	TOTAL
COSTE DIRECTO	9,552.76€	12,416.08€
COSTE INDIRECTO	2863.32€	

Tabla 8 - Resumen de gastos

Estudio previo

1.1. Historia del procesamiento del lenguaje natural.

Para trabajar en el campo del PLN es necesario saber qué es y cómo nace.

El comienzo del procesamiento de lenguaje natural se inicia en la década del 1950. Más concretamente lo inicia Turing en el momento que publica “*Computing Machinery and Intelligence*” (Turing, 1950). Dicho artículo intentaba responder sobre el pensamiento de las máquinas y si estas eran capaces de dicho proceso.

Podemos diferenciar tres fases en la historia de PLN.

1.1.1. Simbólica

Durante el periodo simbólico se establecían unas lógicas y símbolos los cuales daban lugar a un sistema de reglas. Estas inteligencias artificiales incorporaban conocimiento humano para establecer sus conclusiones, además, no aprendían mediante las entradas y salidas sino a partir de la intervención humana.

Algunos trabajos realizados durante esta etapa pueden ser la traducción de más de sesenta frases del ruso al inglés (IBM, 1954) o bien el SHRDLU (Winograd, 1970), un sistema capaz de interactuar con frases en inglés.

1.1.2. Estadística

Con el nacimiento del *machine learning* comienza una nueva etapa en el campo de PLN. Los algoritmos ya trataban de aprender a partir de un cuerpo dado y no se ceñían a unas reglas manuales. Algunos proyectos importantes fueron las máquinas de traducir.

1.1.3. Neuronal

Con la llegada del *Deep Learning* se marca el presente de PLN. Usándose en múltiples campos no solo de la lingüística sino por ejemplo en medicina. Los asistentes de voz que nos acompañan a diario nacen a partir de este proceso de lenguaje profundo y neuronal.

1.2. Introducción al análisis del sentimiento

El análisis del sentimiento ha estado presente en toda nuestra historia. No hablamos del pasado reciente donde podemos usar técnicas de PLN para analizar textos e incluso videos que nos permitan clasificar de forma automática. Nos referimos a épocas en las que era de vital importancia analizar el sentimiento general de un grupo de sujetos determinados. En política por ejemplo, siempre ha sido objeto de estudio el sentimiento general de los posibles votantes. Además hoy en día sigue utilizándose con este fin entre muchos otros.

En los últimos años el interés en el campo del análisis del sentimiento ha ido en aumento (Google, s.f.) reflejado sobre todo en el interés en búsquedas de Google cómo también en el número de estudios que se han realizado.



Ilustración 7 - Evolución de las búsquedas de "Sentiment analysis" a lo largo del tiempo

Diversos estudios se han ido realizando en los últimos años, por ejemplo en el campo de las valoraciones de los productos en internet (Mouthami, 2013), en el campo de la medicina (Deng, 2015) o en el campo de las finanzas. (Mishev, 2020)

apoyándonos en técnicas vistas previamente, más concretamente, en técnicas PLN podemos analizar grandes cantidades de datos en periodos breves de tiempo en comparación al análisis manual.

Estos trabajos se han realizado sobre todo en inglés, donde otros idiomas han sido menos protagonistas en este campo de investigación. Sin embargo podemos encontrar varios trabajos realizados en otros idiomas cómo en el árabe (Magdy, 2021) dónde se realizaron comparación entre distintos métodos de análisis o, por ejemplo, en coreano. (Yang, 2021)

2. Evolución de las tecnologías

Desde el nacimiento de la tecnología PLN se han pasado por diversas etapas en cuanto a la tecnología que se ha ido utilizando.

En 1954 nace la **bolsa de palabras** (Bag of words). Este modelo de representación se basaba en el número de ocurrencias de cada uno de los términos en el corpus. Es un modelo estadístico donde el resultado, teniendo en cuenta los nuevos avances, queda bastante atrasado respecto a sus sucesores.

Con la llegada de las redes neuronales, que simulaban el modelo cognitivo de los seres humanos, llegan distintas formas de operar en PLN.

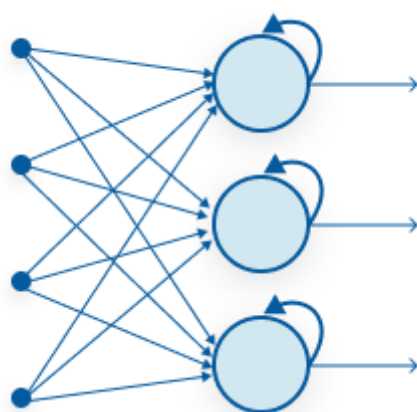
Word Embedding (Word2Vec): En 2013 nace el algoritmo Word2Vec que ya utilizaba el modelo de redes neuronales. Con word2vec las palabras son representaciones numéricas, ya que una

máquina trabaja mejor con números que con conjuntos de letras. Mediante esta técnica cada palabra tenía una característica numérica para compararlas con otras palabras. Los vectores se comparaban para buscar similitudes en sus características y permitirnos ver su similitud o bien sus diferencias. Estos algoritmos se utilizan, por ejemplo, en sugerencias de un teclado móvil.

Dentro del Word embedding existen dos arquitecturas.

- Modelo de palabras continuo (CBOW) que dada múltiples palabras o contexto nos resulta una salida
- Skip-gram continuo, que dada una sola entrada nos resulta múltiples salidas o contexto.

Nacimiento de las redes neuronales recurrentes: Gracias al nacimiento de este tipo de red neuronal podemos hablar de algo parecido a la memoria. Las neuronas daban una salida hacia atrás y no siempre seguía hacia delante



Gracias a esto podía recordar información pasada, retroalimentándose una neurona a sí misma, por ejemplo.

Tenemos dos tipos:

- Redes recurrentes simples
- Redes recurrentes complejas

Ahora vamos a ver un poco en detalles cada uno de estos tipos de redes recurrentes.

Ilustración 9 - Composición neuronal del tipo RNN

Redes recurrentes simples: Posee una memoria muy a corto plazo, se suelen utilizar en el reconocimiento de voz o escritura.

Redes recurrentes complejas: La retroalimentación que obtiene la neurona no solo lo obtiene de sí misma, si no de cualquiera que estuviese a su alrededor. A partir de aquí nace un tipo de red muy popular. Las Long Short-Term Memory (LSTM).

Inventadas en 1997 por Hochreiter y Schmidhuber. Este tipo de red se ha utilizado para el procesamiento de imágenes, reconocimiento de voz y la comunicación. Hay gran variedad de este tipo de red dependiendo de su arquitectura que se han analizado en algunos estudios (Weyrich, 2021)

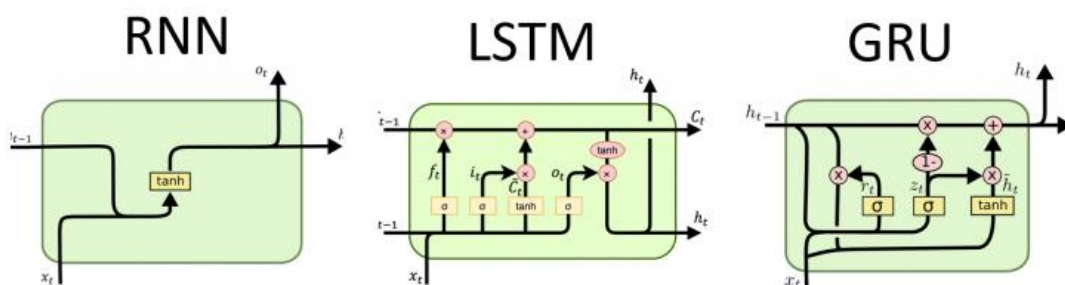


Ilustración 10 - Arquitectura de las redes RNN, LSTM y GRU

La estructura de cada una de las redes varía según su tipo. Podemos observar que las RNN son las más simples y rápidas, sin embargo los resultados no serán tan buenos como con otra de las siguientes arquitecturas. GRU es menos compleja que la del tipo LSTM, lo cual la hace computacionalmente más barata y por lo tanto más rápida. No termina de ser una memoria a largo plazo aunque tampoco se pueda decir que sea de corto plazo, nos encontramos en un punto medio entre las del tipo RNN y LSTM. Finalmente la complejidad del tipo LSTM es mayor, de ahí a que su memoria sea más larga de sus compañeras.

Modelos de atención y encoders-decoders: Gracias al descubrimiento de Levy et al. (Omer Levy, 2015) nos dimos cuenta de que no había una gran diferencia de rendimiento en los traductores entre los modelos clásicos y los nuevos si se realizaba un ajuste correcto. Es por ello por lo que Bahdanau et al. (Bengio, 2016) Basa su modelo predictivo en tomar solo las partes relevantes de la frase de entrada para sacar la frase de salida y no tomar así toda la frase, ya que según el estudio esto produce un cuello de botella.

Otros de los problemas que solucionó fue por ejemplo, en las traducciones, la importancia del orden de las palabras. A la hora de hacer las traducciones se daba una importancia a las palabras dentro de una frase (un peso). Sin embargo esto era engorroso por lo que nacieron los Transformers.

Transformers: Google propone este nuevo modelo de autoaprendizaje basado en la atención y los pesos en la entrada en 2017 y se desprende de las RNN. (Polosukhin, 2017) Este modelo no solo se aplica a PLN, también se usa en otros campos como en la detección de imágenes.

Transformers transforma cada palabra en tres perspectivas distintas (key, query y value). Se calcula la función *softmax* de la siguiente función $\rightarrow (\frac{QK}{\sqrt{dk}})V$ para obtener la puntuación (entre 0 y 1) o importancia de la palabra.

Google ya no utilizaba RNN si no redes de tipo feed-forward (en una sola dirección) las cuales son más clásicas por lo que son más simples.

Modelos pre-entrenados: En la actualidad es común usar un modelo pre-entrenado y afinarlo posteriormente. Un ejemplo de estos modelos pre-entrenados es BERT el cual tiene varias ramificaciones. BERT fue creado por Google en 2018. (Devlin, 2018) Cómo hemos dicho estos modelos deben afinarse para el contexto deseado por el usuario que hace uso de él, haciendo un entrenamiento posterior del mismo modelo.

3. Aporte de este proyecto

Tal y cómo hemos visto, la gran mayoría de estudios se centran en un campo determinado para un análisis concreto. Lo que nosotros pretendemos en este proyecto es utilizar los últimos aportes al campo de PLN. Para ello usaremos Transformers con modelos pre-entrenados, en concreto usaremos BERT como también las librerías de HuggingFace. Además el proyecto también dará soporte al idioma inglés y español. Lo que se pretende aportar es la posibilidad de monitorear en tiempo real las redes sociales haciendo uso de los últimos avances en PLN cómo también cualquier texto dado por un usuario. No solo queremos dar al usuario la respuesta en términos de positividad o negatividad del texto, sino también la agresividad que este transmite, dando así un estudio completo del mensaje.

3.1. Limitaciones y contexto

El presente proyecto se trata de un trabajo de final de grado, lo que supone un límite de tiempo y recursos en comparación con otros proyectos.

En términos de tiempo para este proyecto hemos contado con un total de 4 meses para cumplir todos los objetivos del proyecto, lo cual teniendo en cuenta que el equipo de desarrollo es de una sola persona limita la calidad del software. Para paliar este hecho, se ha ido evaluando los objetivos del proyecto a medida que este ha ido desarrollándose.

Los recursos económicos son parte fundamental para el desarrollo de un proyecto software, ya que este aparte de abrir la posibilidad de tener terceros trabajando en un proyecto, también abre la puerta a funciones avanzadas de las herramientas que se utilizan. Un ejemplo para este proyecto sería el poder adquirir datos de pago para entrenar las redes, utilizar la API premium de Twitter o bien una máquina virtual en la nube (como puede ser de Google o Amazon) para entrenar las redes que hemos desarrollado en este proyecto.

Para poner en contexto el sistema en el que se ha desarrollado el proyecto, se especifica el sistema en el cuadro siguiente.

Hardware	Descripción
Sistema operativo	Windows 10 Pro-64 bits
Procesador	AMD Ryzen 5 1600 Six-Core (12 CPUs) – 3.2 GHz
Memoria RAM	16384 MB
Tarjeta Gráfica	NVIDIA GeForce RTX 2060 SUPER – 8192MB
Discos	SSD M.2 512GB, SSD 124GB, HDD 1024GB

Tabla 9 - Especificaciones hardware del sistema de desarrollo

Datasets y datos de entrenamiento

1.1. Qué es y cuáles son los datos objetivos de este proyecto.

Para el entrenamiento de redes neuronales se usan conjuntos de datos (o en inglés “Datasets”) los cuales están recopilados para un objetivo en concreto.

*Un **conjunto de datos** (conocido también por el anglicismo **dataset**, comúnmente utilizado en algunos países hispanohablantes) es una colección de datos habitualmente tabulada. En el caso de datos tabulados, un conjunto de datos contiene los valores para cada una de las variables organizadas como columnas (Wikipedia, Wikipedia - Conjunto de datos, s.f.)*

Los datos que se han buscado para este proyecto son aquellos que han sido calificados según el significado semántico. En nuestro caso esta clasificación consistía en la agresividad y grado de positividad de una frase concreta.

Tweet/Comentario	Etiqueta
Ya deja de intentar contarle tus problemas a alguien. Entiende A NADIE LE IMPORTAS!!!!	1
"De las peores cosas de la depresión es que no te deja ganas de vivir... Ni de matarte.	1

Tabla 10 - Ejemplo de dataset en español

Tal y como hemos indicado en el apartado de Limitaciones y contexto visto anteriormente, hemos de buscar datasets gratuitos y de licencia libre. Para ello hemos utilizado sitios webs como Kaggle y huggingFace.

Se han utilizado un total de 221446 comentarios/tweets para entrenar los distintos modelos que se han realizado en el proyecto, en la siguiente tabla podemos ver cómo se han distribuido los datos según el modelo que ha sido entrenado.

Idioma	Tipo de análisis	Red social	Nº Tweets/Comentarios
Inglés	Sentimiento	Twitter	99988
		Reddit	37149
	Agresividad	Twitter	31961
		Reddit	37162
Español	Sentimiento	-	8093
	Agresividad	-	7093

Tabla 11 - Distribución de datos en los modelos entrenados

Datasets	Enlace
Evaluación de sentimientos 2018	sem_eval_2018_task_1 · Datasets at Hugging Face
Twitter sentiment analysis	Twitter sentiment analysis Kaggle
Reddit Sentimental analysis Dataset	Twitter and Reddit Sentimental analysis Dataset Kaggle
Ucberkeley-dlab/Measuring-hate-speech	ucberkeley-dlab/measuring-hate-speech · Datasets at Hugging Face
hate_speech_offensive	hate_speech_offensive · Datasets at Hugging Face
tweets_hate_speech_detection	tweets_hate_speech_detection · Datasets at Hugging Face

Tabla 12 - Listado de datasets usados

1.2. Preprocesamiento de los datasets

Los textos que se van a utilizar para entrenar los modelos ya están correctamente clasificados. Sin embargo usar estos datos en bruto no es nada recomendable ya que estaríamos entrenando modelos con información innecesaria, incoherente y sin ningún tipo de información útil. Por ejemplo, cuando añadimos enlaces o menciones a otros usuarios realmente estamos confundiendo el proceso de aprendizaje.

Para limpiar los datos se han seguido los siguientes puntos:

- Convertir el texto en minúsculas
- Eliminar enlaces
- Eliminar las menciones del tipo *@usuario*
- Eliminar los hashtags *#Hashtag*
- Eliminar las palabras *rt* (Usadas por twitter para indicar que es un retweet)
- Eliminación de caracteres especiales
- Eliminación de espacios sobrantes
- Eliminación de números
- Aislar y eliminar puntuaciones excepto ‘?’
- (Solo en inglés) Modificar los ‘t en *not*

Después de realizar todas estas modificaciones en cada uno de los textos del datasets nos quedan datos claros que entran dentro del corpus del modelo pre-entrenado que se utiliza para afinar posteriormente con estos datos.

Texto en bruto	Procesado
(: !!!!! - so i wrote something last week. and i got a call from someone in the new york office... http://tumblr.com/xcn21w6o7	so i wrote something last week and i got a call from someone in the new york office
@ange_black @sween I call dibs on the Voltron arm. No the leg. Wait. Where are my manners? @baileygenine,.. http://tr.im/oERy	i call dibs on the voltron arm no the leg wait where are my manners ?
Están viendo que se le olvidaron tomarse los antidepresivos al muchacho y me ponen canciones que me recuerdan a ella #AsiNoSePuede	están viendo que se le olvidaron tomarse los antidepresivos al muchacho y me ponen canciones que me recuerdan a ella

Tabla 13 - Ejemplo de procesamiento de texto

Entrenamiento de modelos

1. Elección de modelos pre-entrenados y ajuste fino

Para el entrenamiento de los modelos que se van a realizar se va a utilizar un modelo pre-entrenado como ya hemos indicado en el apartado Evolución de las tecnologías. Se ha seleccionado BERT (Toutanova, 2018) como modelo.

1.1. Inglés

Para el inglés se ha seleccionado el modelo *Bert base uncased* que se trata del modelo BERT basado en transformadores.

(Del inglés): El modelo BERT se preentrenó en BookCorpus, un conjunto de datos formado por 11.038 libros inéditos y Wikipedia en inglés (excluyendo listas, tablas y cabeceras). (Face, s.f.)

Además se ha seleccionado la versión uncased, en la que el texto se procesa en minúsculas y sin puntuaciones en las palabras, es decir, si tenemos por ejemplo la palabra *Cry* y *cry*, estas se procesarán como *cry* en cualquier caso.

1.2. Español

Para el español se ha utilizado el modelo *dccuchile/bert-base-spanish-wwm-uncased* de Hugging Face que es un modelo parecido al base de Bert. Este modelo está diseñado en español

2. Funcionamiento de Bert

Bert se basa en Transformers aprendiendo así las relaciones del contexto dentro de un texto. Se basa en dos mecanismos, MLM y NSP.

2.1. Técnica MLM

Esta técnica permite que un modelo sea entrenado de la siguiente manera. Se toma el 15% de las palabras de entrada y se reemplaza con una máscara. El modelo ahora con la entrada restante intentará predecir el valor original del enmascarado según el contexto.

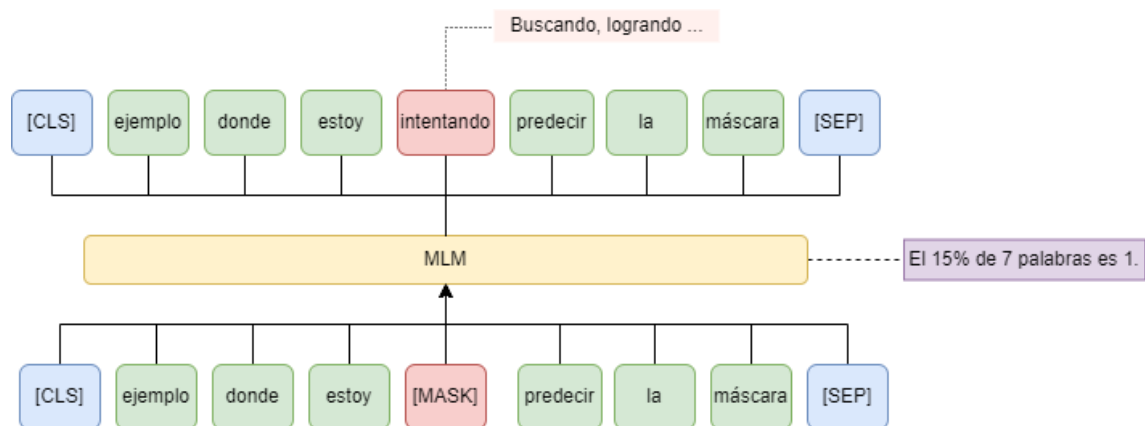


Ilustración 11 –Funcionamiento MLM

2.2. Técnica NSP

Durante este proceso se le ofrece al modelo un par de frases, en la mitad de los casos la segunda frase es la siguiente a la primera y en el resto no, de esta manera el modelo debe predecir cuándo esto está ocurriendo. Para ayudar al modelo se le añade los tokens de inicio de frase a la primera frase y el token de fin de frase en ambas. (Horev, 2018)

Modelos entrenados

1.1. Subdivisión de tareas

Nuestro problema se centra en dos pilares fundamentales.

- Idioma
- Red social

Para ello debemos tener en cuenta que vamos a tener más facilidades en el idioma inglés y en las redes sociales más populares (siendo Twitter la mayor de ellas) tendremos más facilidad de encontrar datos.

Después de la búsqueda de datos llegamos a la conclusión que lo más óptimo es realizar 6 modelos distintos los cuales estarán especializados en la misma tarea pero un idioma diferente.

Es por ello por lo que para el español se han entrenado dos modelos distintos, uno especializado en el análisis de sentimiento y otro en la agresividad, en el caso del inglés se hecho lo mismo pero diferenciando las redes sociales de Twitter y Reddit, resultando en cuatro modelos distintos.

1.2. Primeras pruebas

Partiendo de la base de que se ha tenido que realizar un estudio completo de como entrenar este tipo de modelos, se inicia esta fase de pruebas con distintos datasets y otros tipos de modelos.

Los datos que se utilizaron fueron Sentiment140 (Kaggle, s.f.) y dos modelos pre-entrenados distintos a bert que son más ligeros que el primero: *twitter-roberta-base-sentiment* (Barbieri, 2020) y *distilbert-base-uncased* (Wolf, 2019)

Las primeras pruebas se han enfocado en conocer el proceso de entrenamiento de los modelos. Después de probar varios modelos se llegó al proceso de preprocesamiento de datos que mejor se complementaba con el tipo de entrenamiento que se indica en Preprocesamiento de los datasets.

Se crearon 17 tipos de modelos combinando distintas redes y combinaciones de capas.

Las capas que más han sido utilizadas son las siguientes.

- Capa de densidad (Oculta): (Keras, keras API - Dense layer, s.f.): Se trata de una capa oculta de la red neuronal, podemos darle el número de neuronas que queramos aunque debe ir en decremento. Tiene distintos tipos de activación pero en nuestro caso usamos RELU. Este tipo de activación se usa sobre todo en problemas de clasificación de imágenes, en nuestro caso el problema es distinto al mencionado pero es la activación que mejor ha funcionado (sólo en las capas ocultas). Esto desactivará la neurona si recibe un valor menor a 0.
- Capa de abandono (Keras, Keras - Dropout layer, s.f.): Usada para reducir el sobreajuste. Esta capa pone aleatoriamente las entradas a 0, en la fase de entrenamiento, con una frecuencia dada.

- Capa de densidad (Salida): Funciona de la misma manera a la ya mencionada capa de densidad pero en este caso la usamos como la última capa, esta se prueba con activación *softmax* y *sigmoid*.

(Del inglés): La función sigmoide se utiliza para la regresión logística de dos clases, mientras que la función softmax se utiliza para la regresión logística multiclase (veritessa, s.f.)

En la fase de entrenamiento también utilizamos optimizadores para reducir el error en la modificación de pesos al actualizar la red (paso esencial en el algoritmo de *Backpropagation*).

Los optimizadores que han dado mejores resultados son los siguientes:

- SGD: Limita a una sola observación el cálculo del gradiente, cuando normalmente se toman varias observaciones o todas. Supone menor tiempo para el cálculo.
- Adam: Combina dos metodologías (*Momentum* y *RMSProp*). Toma las actualizaciones anteriores del gradiente y mantiene tasas diferentes para el aprendizaje.

(InteractiveChaos, InteractiveChaos - Stochastic Gradient Descent) (Velasco, 2020)
(InteractiveChaos, Adam - InteractiveChaos)

1.3. Entrenamiento de los modelos definitivos

Tras la realización de pruebas con los primeros modelos se llega a la siguiente conclusión respecto a los hiper parámetros para el entrenamiento.

	Sentimiento			Agresividad		
	Twitter	Reddit	Español	Twitter	Reddit	Español
Longitud de secuencia	150	256	150	150	256	150
Tamaño del batch	12	12	18	16	12	18
% dataset entreno	75 %					
N.º capas densidad y abandono	Densidad: 5 Abandono: 2					
N.º Salidas	1	3	2	1	2	2
Optimizador	SGD					

Tabla 14 - Hiper parámetros de los modelos

La longitud de secuencia tiene relación con los conjuntos de datos, los comentarios en Reddit suelen ser más largos de ahí que la longitud sea mayor que en los conjuntos de twitter, al igual que el batch.

Los pasos a seguir para entrenar los modelos han sido los mismos para los seis modelos distintos. Para enfocar el entrenamiento seguido podemos ver los siguientes esquemas.

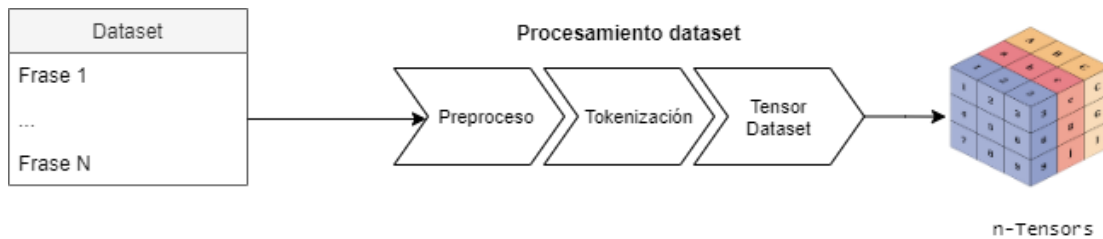


Ilustración 12 - Procesamiento datasets

Cómo se puede apreciar en el diagrama anterior, primero se carga el dataset objetivo del modelo a entrenar usando el paquete pandas para el análisis de datos. Una vez cargado se preprocesa los datos tal y cómo hemos visto en el apartado Preprocesamiento de los datasets e iniciamos la etapa de tokenización.

Las etiquetas por otra parte se procesan para que sean de tipo binario y sea coherente con el problema, además estas se procesan con el método one-hot encoding, donde se pasan a un array del tamaño del número de posibilidades de clasificación, dejando en “1” la posición a la que hace referencia la etiqueta.

De esta manera podemos hacer que el modelo entienda la clasificación correspondiente.

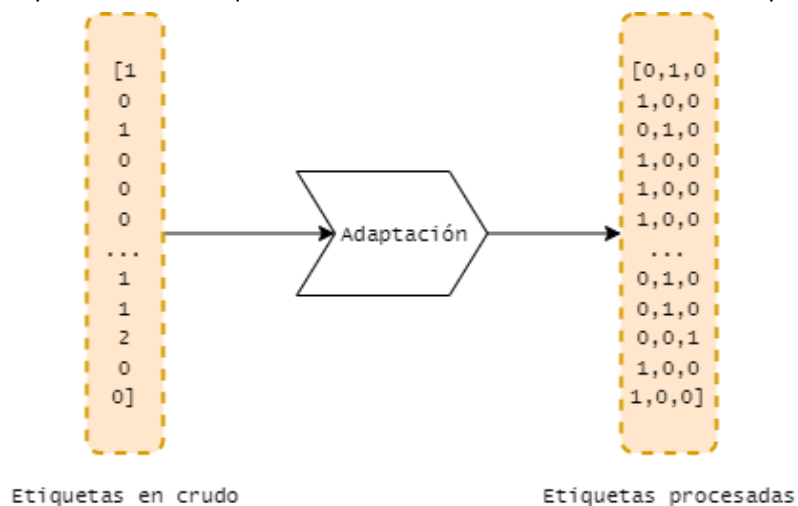


Ilustración 13 - Procesamiento de etiquetas

El proceso de tokenización es algo más complejo. En nuestro caso la tokenización se realizará por palabras completas. Estas palabras se cambiarán por valores numéricos junto a una máscara de atención, la cual indicará que posición del array que se va a formar debe tener en cuenta la red neuronal. ¿Por qué tokenizar? Los sistemas informáticos entienden mejor los números que el lenguaje humano, es por ello por lo que cada palabra lleva asociada un valor decimal. Para realizar la tokenización se ha utilizado el tokenizador de Bert añadiendo tokens especiales como los tokens indicadores del inicio de una frase o su final. Todos los tokens son del tamaño de la secuencia, de ahí que la máscara de atención sea importante para poder procesar la información de entrada, de modo que si un texto no alcanza esa longitud se añade un *padding* (una serie de ceros). Es importante que se utilice el apropiado para el modelo que se va a usar para afinar.

Para ilustrar mejor cómo funciona la tokenización podemos utilizar el diagrama siguiente para ver el ejemplo de tokenización de un conjunto de frases.

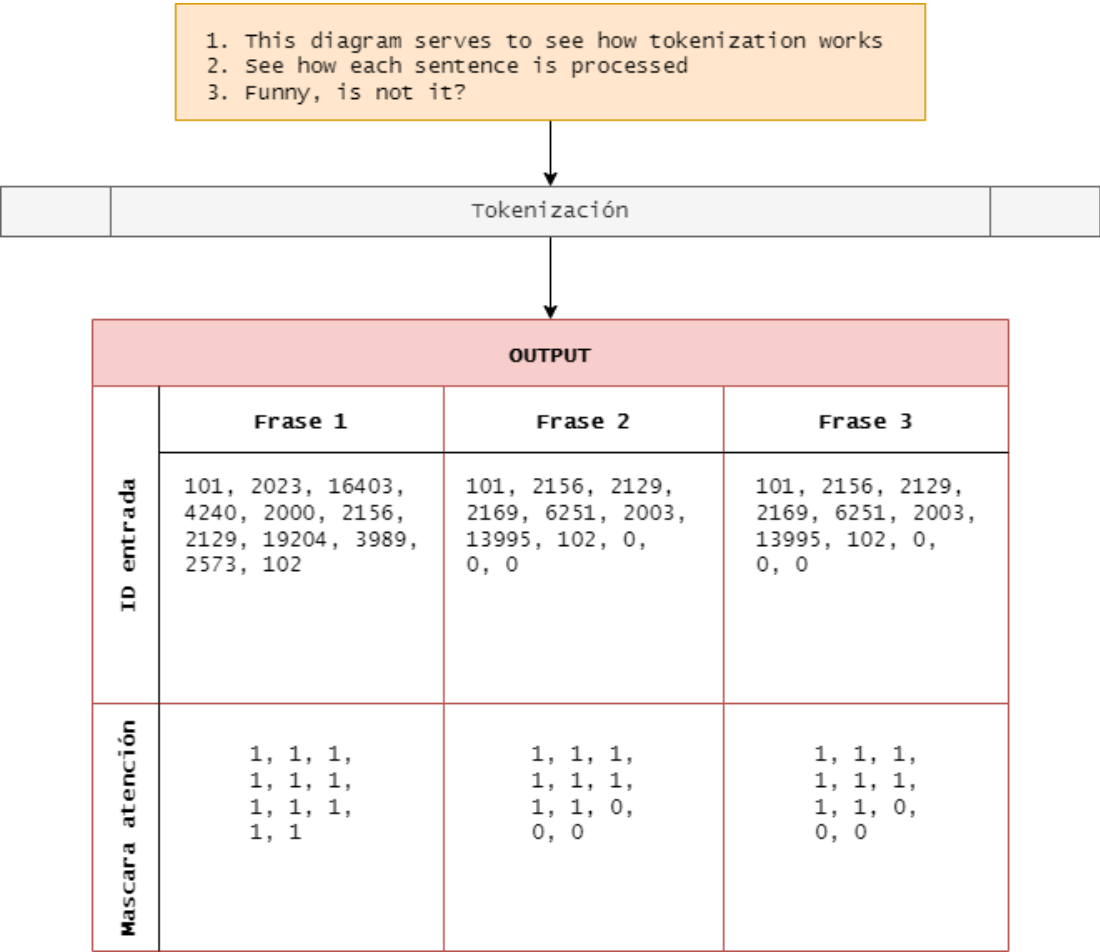


Ilustración 14 - Proceso de tokenización

Antes de entrenar el modelo debemos seleccionar el optimizador además de otros aspectos como puntos de control y paradas tempranas.

El optimizador que se ha utilizado es SGD centrado en mejorar la precisión del algoritmo para clasificación binaria (excepto el modelo de Reddit de sentimiento el cuál se adapta posteriormente).

Los puntos de control sirven para que el modelo vaya guardando sus pesos cada cierta etapa para así poder tener la red neuronal en otros puntos del entrenamiento y poder estudiarlo posteriormente.

Los Early Stopping se utilizan para que el modelo pare de entrenar si la métrica a seguir está empeorando desde hace ciertas etapas para así ahorrar recursos y ajustar el modelo según los resultados obtenidos. También si no utilizamos los puntos de control a la vez, sirve para tener el mejor modelo disponible que ha sido entrenado (sin entrar en sobreajuste). Después de configurar tanto el optimizador como el early stopping y los puntos de control, podemos

empezar a entrenar el modelo pasándole el conjunto de entrenamiento y el conjunto de validación.

El algoritmo de *backpropagation* funciona de la siguiente manera tal y cómo está indicado en el siguiente diagrama.

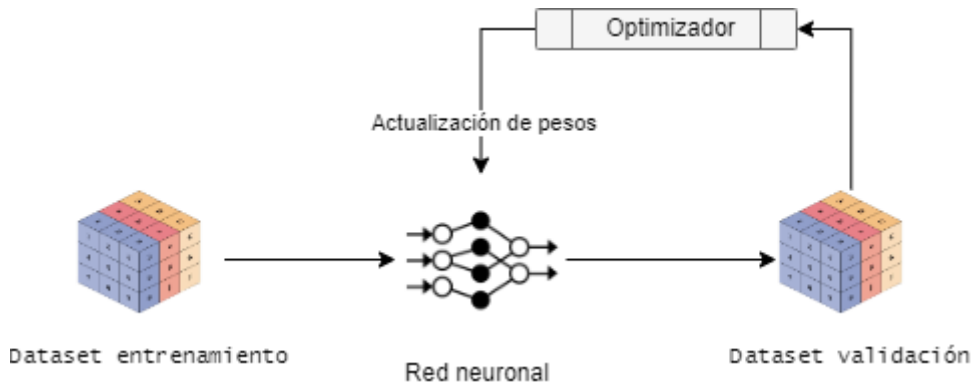


Ilustración 15 - Algoritmo de *backpropagation*

El algoritmo funciona de la siguiente manera:

1. El conjunto de entrenamiento pasa por la red neuronal
2. El conjunto de validación pasa por la red neuronal después de que esta haya sido actualizada en el paso anterior
3. Se calcula el coste o el error y se pasa por el optimizador
4. El optimizador calcula los nuevos pesos y se actualiza la red
5. Si quedan más etapas se repite el entrenamiento desde el paso 1. Si no se devuelve la red.

Las arquitecturas de las redes entrenadas son algo diferentes por la naturaleza de los datos.

En el caso de las redes que funcionan con la salida de *sigmoid*, se opera de forma que tengamos los datos igual a *softmax*. Al igual que el modelo de Reddit que nos da tres posibilidades (positivo, negativo y neutral), lo operamos de manera que solo tenemos en cuenta las dos etiquetas que usan el resto de los modelos (positivo y negativo)

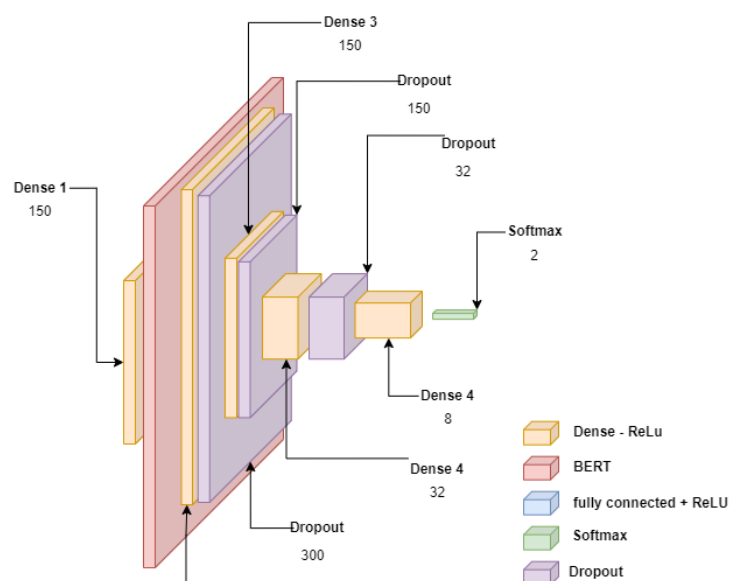


Ilustración 16 - Red neuronal tipo entrenada

1.4. Evaluación de los modelos resultantes

Para evaluar los modelos se han tomado distintas métricas.

- Resultados en las predicciones del dataset de pruebas. Se toman dos valores de esta métrica:
 - Precisión: Se trata del número de clasificaciones correctas entre el número total de clasificaciones.
 - Pérdida: Indica si la predicción del modelo entrenado ha sido buena o no. Mientras más cerca de cero mejor resultado.
- Evolución de resultados en el entrenamiento. Evolución de la precisión y pérdida durante la fase de entrenamiento en forma de gráfica.
- Matriz de confusión: Muestra cómo ha funcionado el algoritmo con una matriz que presenta tanto los falsos positivos como verdaderos positivos de cada etiqueta.
- Curvas ROC: Evalúa los modelos de clasificación mostrando una curva que muestra cómo distingue el modelo dos etiquetas distintas. Mientras más arriba a la izquierda mejores resultados.

Todas las métricas se han comprobado para tener un modelo lo menos sobre-ajustado y mejor entrenado posible.

1.4.1. Modelos sentimiento

1.4.1.1. Twitter

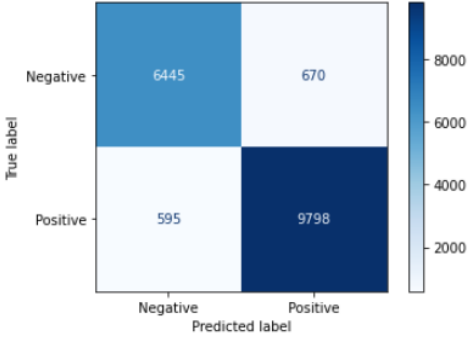
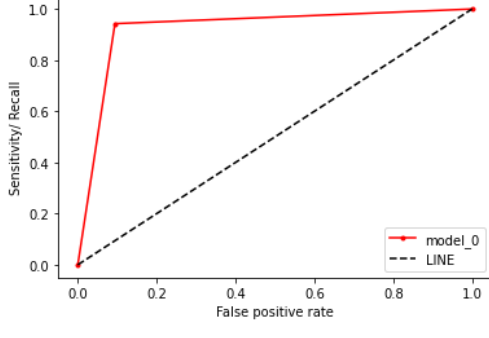
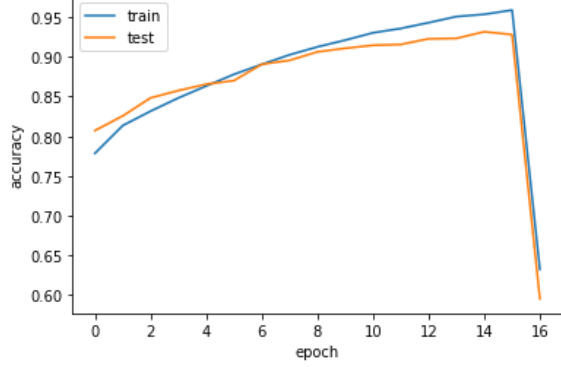
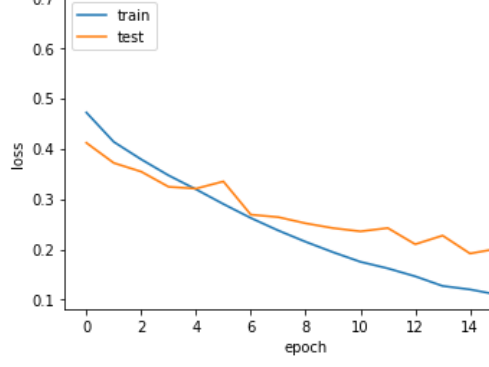
Matriz confusión	Curva ROC
	
Evolución de la precisión	Evolución de la pérdida
	
Etapas de parada	15
Precisión dataset pruebas	0.9277
Pérdida dataset pruebas	0.2048

Tabla 15 - Resultados modelo twitter de sentimiento

1.4.1.2. Reddit

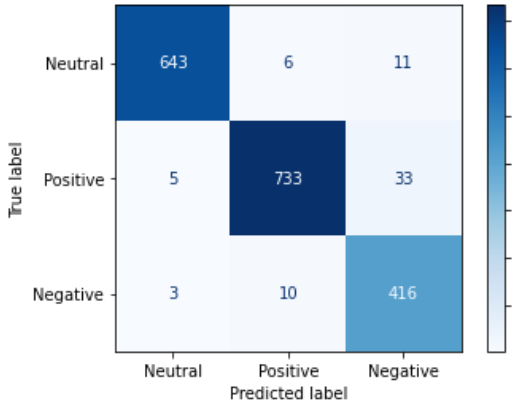
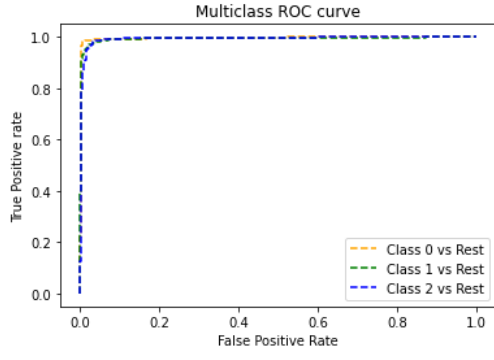
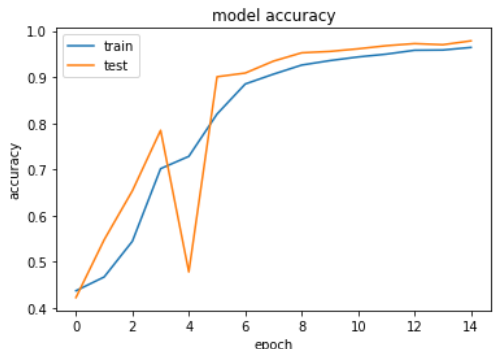
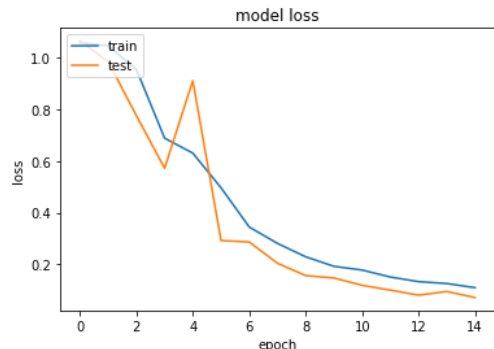
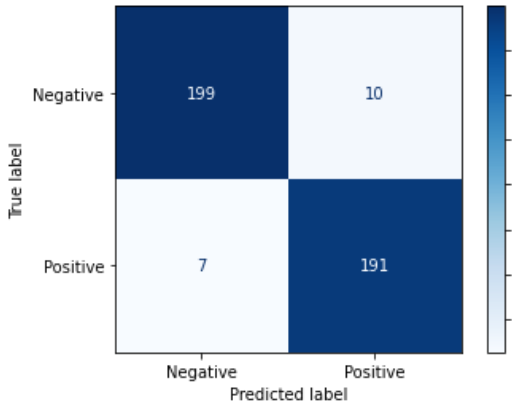
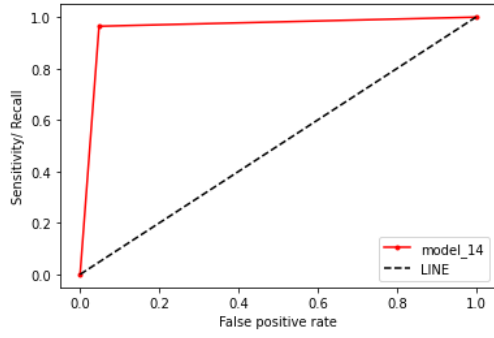
Matriz confusión		Curva ROC	
 <p>True label</p> <p>Predicted label</p> <p>Neutral Positive Negative</p> <p>Neutral Positive Negative</p>		 <p>Multiclass ROC curve</p> <p>True Positive rate</p> <p>False Positive Rate</p> <p>Class 0 vs Rest Class 1 vs Rest Class 2 vs Rest</p>	
Evolución de la precisión		Evolución de la pérdida	
 <p>model accuracy</p> <p>accuracy</p> <p>epoch</p> <p>train test</p>		 <p>model loss</p> <p>loss</p> <p>epoch</p> <p>train test</p>	
Etapa de parada		12	
Precisión dataset pruebas		0.9634	
Pérdida dataset pruebas		0.1158	

Tabla 16 - Resultados modelo reddit de sentimiento

1.4.1.3. Español

Matriz confusión		Curva ROC	
 <p>True label</p> <p>Predicted label</p> <p>Negative Positive</p> <p>Negative Positive</p>		 <p>Sensitivity/ Recall</p> <p>False positive rate</p> <p>model_14 LINE</p>	
Evolución de la precisión		Evolución de la pérdida	

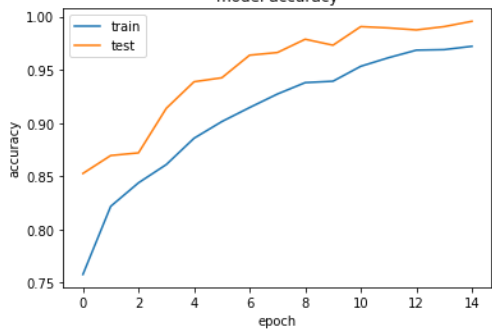
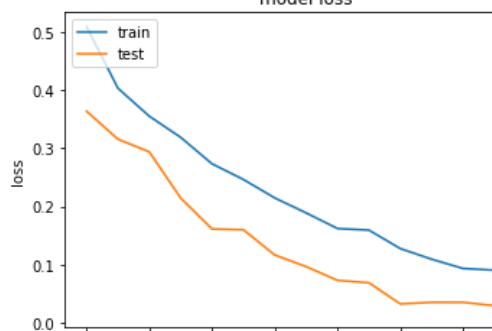
	
Etapas de parada	16
Precisión dataset pruebas	0.9582
Pérdida dataset pruebas	0.1092

Tabla 17 - Resultados modelo español de sentimiento

1.4.2. Modelos agresividad

1.4.2.1. Twitter

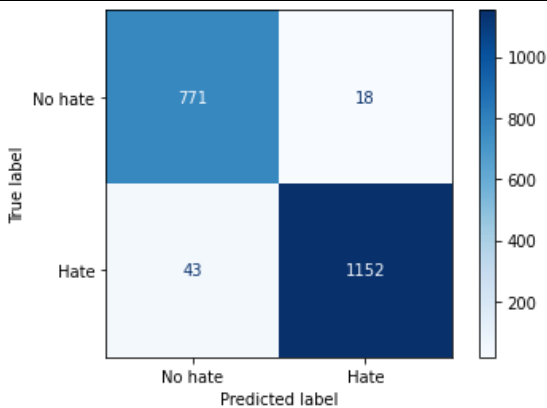
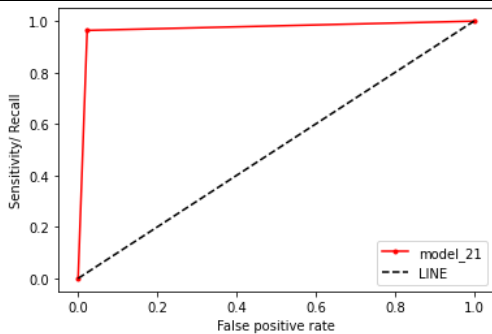
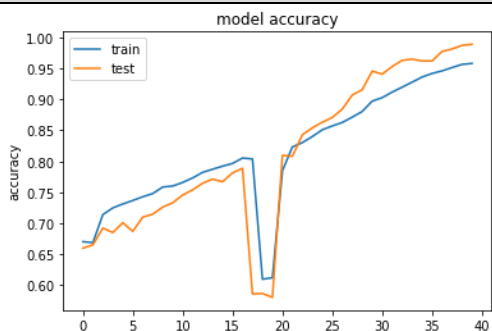
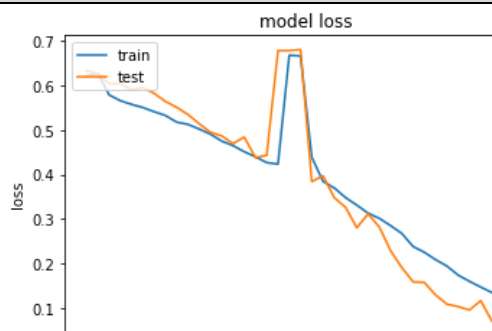
Matriz confusión 	Curva ROC 
Evolución de la precisión 	Evolución de la pérdida 
Etapas de parada	33
Precisión dataset pruebas	0.9692
Pérdida dataset pruebas	0.1038

Tabla 18 - Resultados modelo twitter de agresividad

1.4.2.2. Reddit

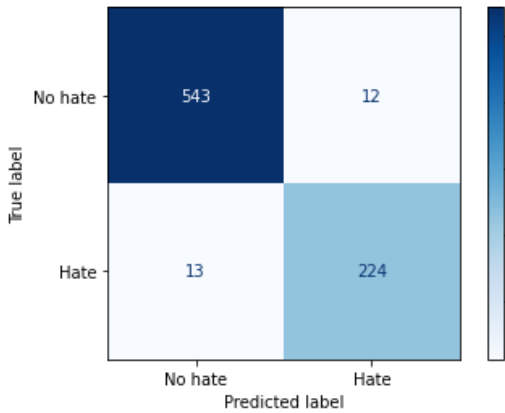
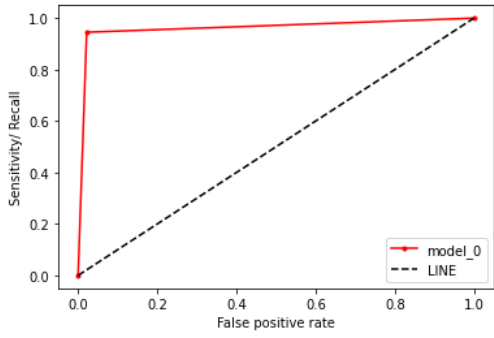
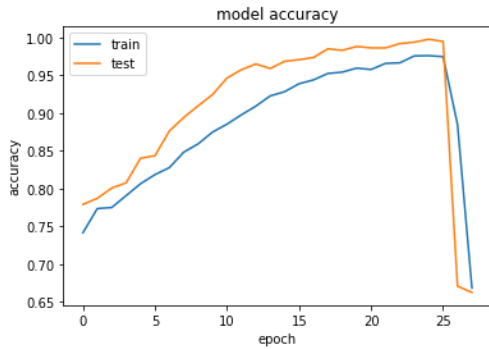
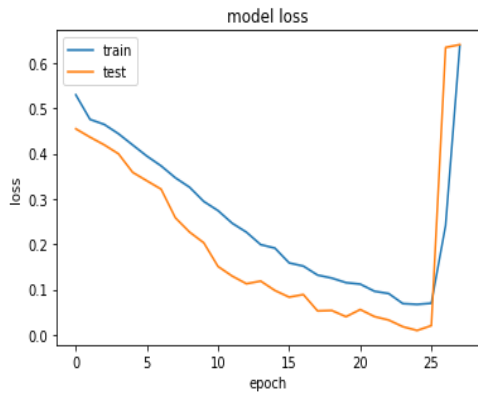
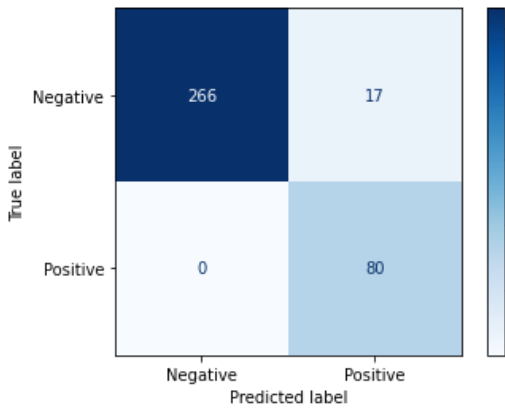
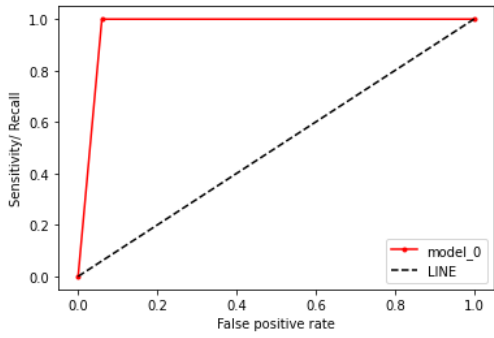
Matriz confusión	Curva ROC
	
Evolución de la precisión	Evolución de la pérdida
	
Etapas de parada	12
Precisión dataset pruebas	0.9684
Pérdida dataset pruebas	0.1129

Tabla 19 - Resultados modelo reddit de agresividad

1.4.2.3. Español

Matriz confusión	Curva ROC
	

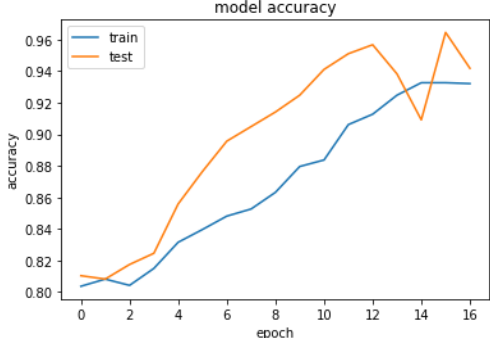
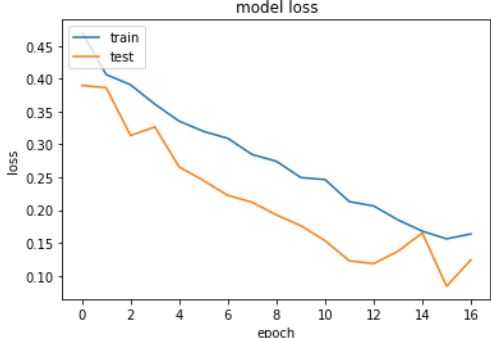
Evolución de la precisión	Evolución de la pérdida
	
Etapas de parada	18
Precisión dataset pruebas	0.9531
Pérdida dataset pruebas	0.1145

Tabla 20 - Resultados modelo español de agresividad

1.4.3. Conclusiones

Como podemos observar se han conseguido modelos que cumplen con los objetivos propuestos. Estos modelos han sido entrenados con los datos disponibles de forma gratuita y con un entrenamiento y cambios en el modelo basados en la prueba y el error, que a su vez han formado parte del aprendizaje personal sobre la materia que se está aplicando en este proyecto.

Es por ello por lo que, lejos de ser perfectos, son modelos totalmente válidos para la consecución de los objetivos de investigación.

Motivación del desarrollo

Las redes sociales son una mina de opinión. El gran problema es la gran cantidad de datos por segundo que se suben a estas redes. Según la revista Forbes, 456.000 Tweets son enviados cada minuto. (Marr, 2018). Procesar esta cantidad de datos de forma manual es tediosa y una actividad prácticamente imposible.

Aprovechar los modelos que han sido entrenados en el capítulo anterior puede ser una manera rápida y automática de procesar todos los datos de interés que necesitamos, aunque para ello necesitamos una aplicación que de soporte a este tipo de tareas.

La mayoría de las aplicaciones destinadas a este cometido están alojadas en la nube, lo cual no sería un problema si nuestra tarea fuese sencilla (procesar pocos datos). En nuestro caso queremos hacer estas tareas de forma masiva pero de manera gratuita, lo cual se encarece bastante si queremos procesar esto de forma alojada. Por ello diseñaremos una aplicación de escritorio para aprovechar la potencia local y además que sea totalmente gratuita.

Tecnologías

Las tecnologías a utilizar para este proyecto son las siguientes:

- Python
- Interfaz gráfica Tkinter
- Modelos previamente entrenados
- GitHub
- Mockup Classic

La elección de Python se debe a que es el mejor lenguaje para este tipo de tareas por las bibliotecas que contiene, además de que todos los modelos que hemos entrenado han sido en Python.

La elección de Tkinter es una manera sencilla de dar al usuario una interfaz gráfica sin demasiada complejidad en el desarrollo. Las limitaciones que tiene esta herramienta han sido subsanadas con pequeños arreglos usando otro tipo de librerías.

El siguiente paso que dar se trata de los mockups o bien los diseños de la aplicación que hemos desarrollado.

Mockups

1. Importancia del diseño

El diseño software es una etapa fundamental en el desarrollo de un proyecto. Esta etapa nos permite plasmar la solución a la que se quiere llegar según los objetivos del proyecto, además de contemplar los requisitos del sistema.

Documentar la solución que se ha realizado con el diseño nos facilita el desarrollo pues ya tenemos un producto visual al que llegar.

Esta etapa nos ahorra bastantes horas de trabajo posterior además de poder detectar a tiempo posibles incongruencias en la solución planteada para el problema propuesto.

El diseño no es cerrado. Es decir, a medida que se va desarrollando y avanzando el proyecto, este diseño puede sufrir cambios, ya sea por la planificación o bien por el propio desarrollo software en el cual pueden surgir problemas y contratiempos.

2. Herramienta utilizada

La herramienta utilizada para ha sido MockPlus usando la licencia gratuita. La suscripción gratuita ofrece lo siguiente:

Usuarios	10
Proyectos	10
Paginas	100
Número de días disponible	14

Tabla 21 - Suscripción gratuita MockPlus

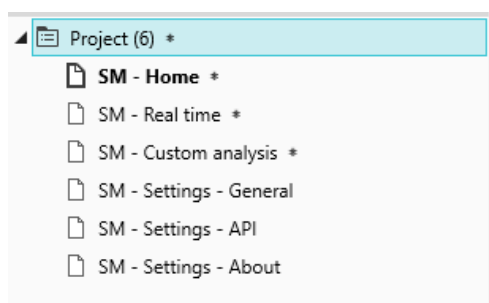


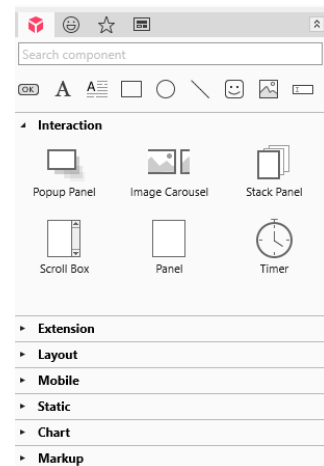
Ilustración 17 - MockPlus: Vista de proyecto

La herramienta tiene una vista con las distintas partes de la aplicación a diseñar, permitiendo conectar cada una de sus partes.

Podemos iniciar la presentación e interactuar con ella moviéndonos por cada una de sus partes.

Además ofrece una biblioteca propia de iconos y figuras para añadir. Las cuales ofrece distintas opciones.

Son elementos en los que podemos interactuar una vez que iniciemos la presentación. Todas estas opciones están disponibles en la suscripción gratuita



3. Ilustraciones y explicación del diseño

3.1. Ilustraciones y explicación

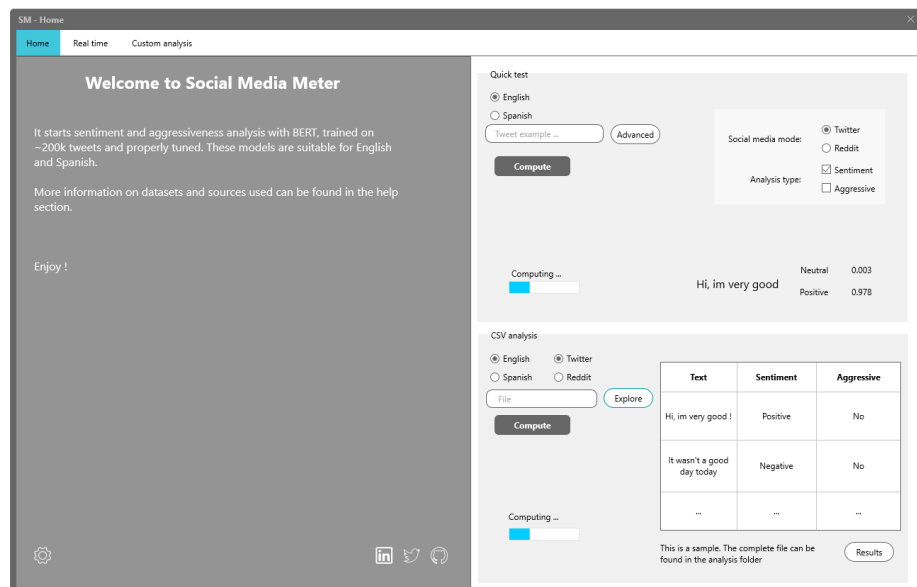


Ilustración 19 - SM-Meter Home

El apartado home está compuesto por 4 partes diferenciadas.

- El texto de bienvenida. Contiene el botón de opciones y otros tres botones haciendo referencia a las redes sociales.
- Apartado superior derecha. Sirve para realizar un análisis rápido. Podemos seleccionar si queremos hacer el análisis basado en Twitter o Reddit y el idioma. Cuando ejecutemos el análisis se mostrará una barra de progreso y finalmente los resultados abajo a la derecha.
- Apartado inferior derecha. Sirve para realizar un análisis de un fichero CSV. Contiene un botón para buscar el archivo en el propio equipo y dos botones para seleccionar el idioma (Finalmente los botones de red social se han suprimido y sustituidos por una selección automática dependiendo de la longitud del texto). Además se ha añadido un input para seleccionar la columna a analizar del fichero. Una vez finalizado el análisis aparecerá una pequeña tabla con los resultados. Si se quiere ver en profundidad podemos darle al botón de *Results* para mostrar todos los resultados cómo en la siguiente ilustración.

- Menú superior. Contiene las tres ventanas principales. El apartado principal Home, el análisis en tiempo real y el análisis personalizado.

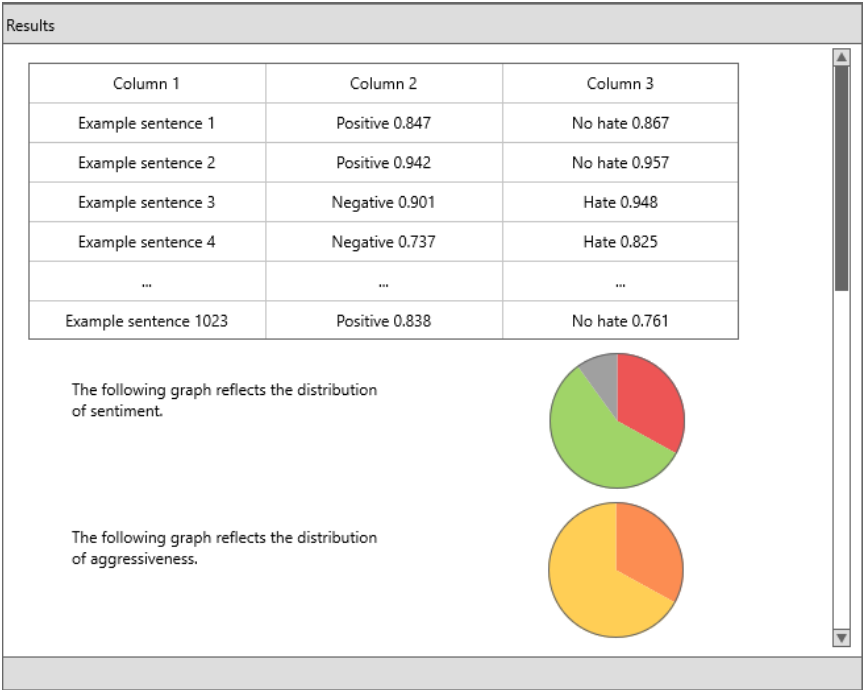


Ilustración 20 - SM-Meter: Resultados de un análisis de CSV

Este apartado muestra los resultados de un análisis de CSV. Se muestran los resultados en tres columnas. La primera trata del texto y las dos siguientes los resultados en sentimiento y agresividad. Los resultados son acompañados de dos gráficos que muestran la distribución total.

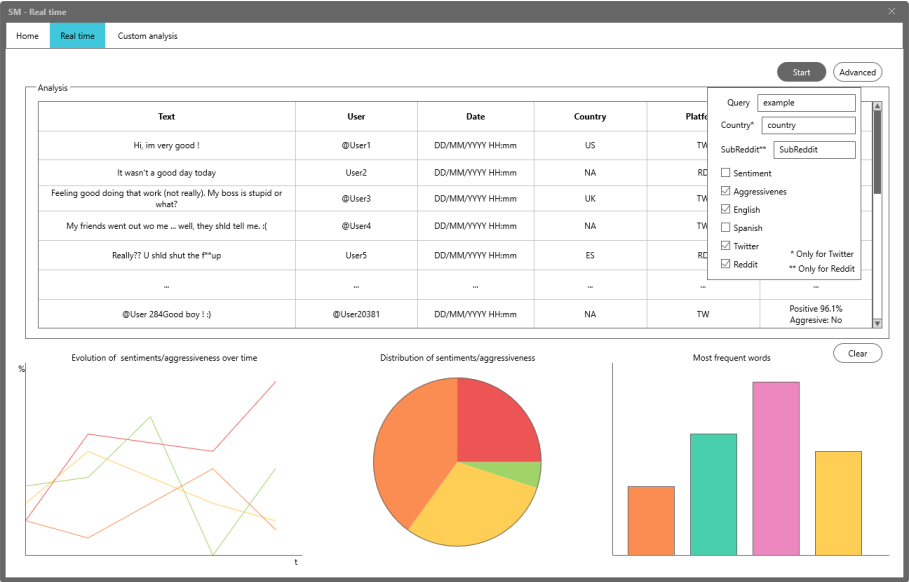


Ilustración 21 - SM-Meter Análisis en tiempo real

Este apartado muestra los resultados en tiempo real de un análisis. Tiene las siguientes partes.

- Opciones avanzadas: Contiene la query a buscar, el país (Solo para Twitter), subreddit (Solo para Reddit) y las distintas opciones de idioma, modo y red social.
- Apartado de los resultados: Contiene el comentario, el usuario junto a la fecha, el país y los resultados en sentimiento y/o agresividad
- Gráficos inferiores: Contiene los gráficos de la evolución en tiempo real del sentimiento general, la distribución de los resultados y las palabras más frecuentes.

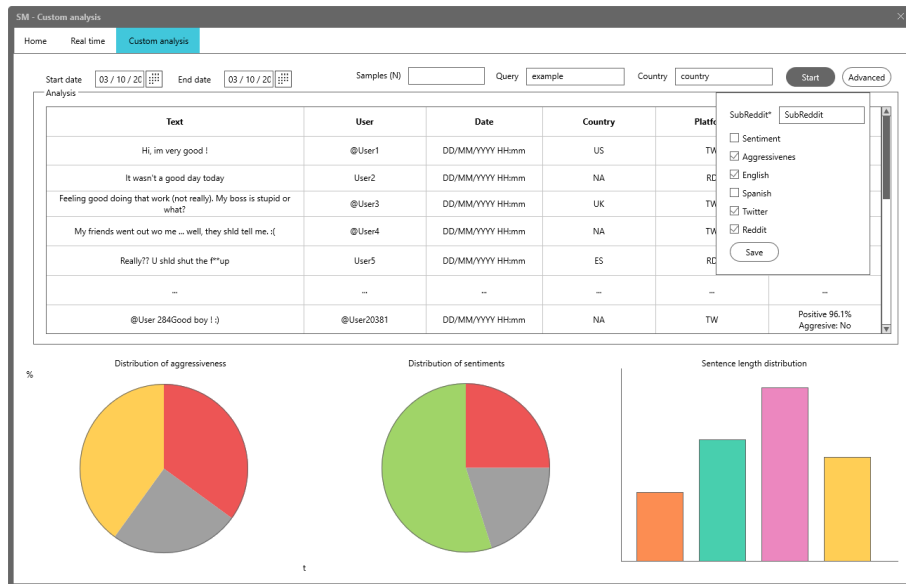


Ilustración 22 - SM-Meter Análisis personalizado

Este apartado muestra los resultados personalizados de un análisis. Tiene las siguientes partes.

- Número de comentarios: Número de comentarios que se van a tomar para análisis
- Query: Palabra a buscar en la red social
- Country: País a filtrar (Solo en Twitter)
- Opciones avanzadas: Subreddit (Solo en Reddit) y las opciones de idioma, modo y red social.

El apartado general de opciones contiene los seis modelos y su estado, es decir, si estos están activos o no.

Tenemos en cada uno de los modelos una serie de botones para cambiar su estado.

Además como en las siguientes ilustraciones se tiene un menú de las tres opciones del menú. General, API y About ...

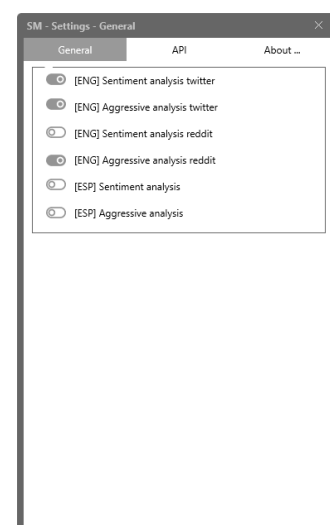


Ilustración 23 - SM-Meter opciones: General

El apartado general de opciones contiene los seis modelos y su estado, es decir, si estos están activos o no.

El apartado de APIs contiene dos apartados importantes. La API de twitter y la API de Reddit por si el usuario quisiera cambiarlas por las propias, ya que estas por defecto son gratuitas.

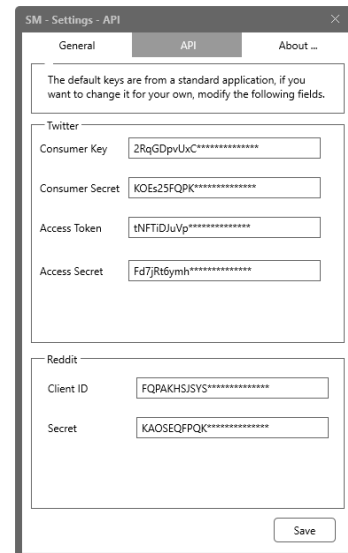
The screenshot shows the 'SM - Settings - API' window. It has three tabs: 'General', 'API', and 'About ...'. The 'API' tab is selected. Inside, there's a note: 'The default keys are from a standard application, if you want to change it for your own, modify the following fields.' Below this, there are two sections: 'Twitter' and 'Reddit'. The 'Twitter' section has four fields: 'Consumer Key' (2RqGDpvUxC*****), 'Consumer Secret' (KOE25FQPK*****), 'Access Token' (tNFTIDJuVp*****), and 'Access Secret' (Fd7jRt6ymly*****). The 'Reddit' section has two fields: 'Client ID' (FQPAKHSJSYS*****), and 'Secret' (KAOSEQFPQK*****). A 'Save' button is at the bottom right.

Ilustración 24 - SM-Meter opciones: API

El último apartado se trata de la información de la aplicación que contiene solamente la versión y la fecha.

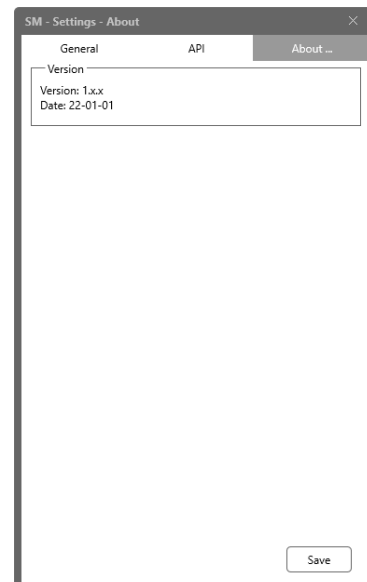
The screenshot shows the 'SM - Settings - About' window. It has three tabs: 'General', 'API', and 'About ...'. The 'About ...' tab is selected. Inside, there's a 'Version' section with two lines of text: 'Version: 1.x.x' and 'Date: 22-01-01'. A 'Save' button is at the bottom right.

Ilustración 25 - SM-Meter opciones: About

Desarrollo

1. Proceso de desarrollo

El proceso de desarrollo se ha dividido en dos partes. Primero se ha realizado toda la interfaz gráfica con Tkinter y luego se ha realizado toda la lógica de backend.

La parte de backend se ha realizado usando multihilos para así poder realizar multiples tareas a la vez. Si el usuario está realizando un análisis, puede comenzar otro nuevo en cualquier momento.

La parte de backend también está mejor soportada si se utiliza con una GPU en lugar de CPU.

Por otro lado el acoplamiento durante el desarrollo ha intentado que sea mínimo. La aplicación consta de cuatro archivos.

1.1. Distribución de responsabilidades

En el siguiente apartado se explica la responsabilidad de los distintos archivos que componen la aplicación.

1.1.1. Main.py

1.1.1.1. Explicación

El archivo main contiene la lógica general y de arranque de la aplicación. Controla la interfaz del usuario y conecta todas las ventanas disponibles. Se encarga también de controlar las opciones de la aplicación y de iniciar los hilos.

1.1.1.2. Métodos

El archivo main contiene la clase app que contiene la lógica de la aplicación. Tiene seis métodos principales los cuales son las distintas pestañas de los menús. Dentro de esta clase también se encuentran los métodos para modificar la configuración o bien aquellos para llamar al método que inicia o para los modelos.

Fuera de la clase están los métodos destinados a crear los gráficos tanto en el análisis en tiempo real, análisis personalizado o el análisis de un fichero CSV.

Por último también contiene el método para abrir el navegador si se clica en los botones de redes como también aquellos que crean las tablas en los resultados.

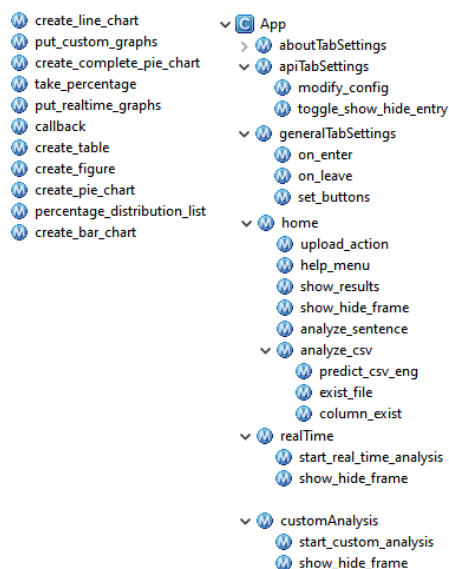


Ilustración 26 - SM-Meter: Métodos en el archivo main.py

1.1.2. ApiSupport

1.1.2.1. Explicación

El archivo apiSupport contiene la lógica de las llamadas a la API de Twitter y Reddit. Se encarga de todo lo que concierne a los distintos análisis que se realizan. Por último también contiene la lógica para exportar los resultados.

1.1.2.2. Métodos

En este fichero se tiene una clase llamada IDPrinter, el cual es el encargado de hacer el streaming de comentarios de la plataforma Twitter.

A parte de esta clase también contiene todos los métodos encargados de los análisis tanto en tiempo real cómo personalizados.

Los métodos más importantes son tanto el encargado de tomar las credenciales de los archivos de configuración cómo de insertar comentarios en las tablas de análisis. También contiene los métodos para hacer llamada al último archivo encargado de los modelos neuronales.

Este archivo al contener variables globales también contiene métodos para actualizar estas cuando son necesarias.

En el siguiente apartado (1.2) puede verse el flujo de llamadas entre archivos.

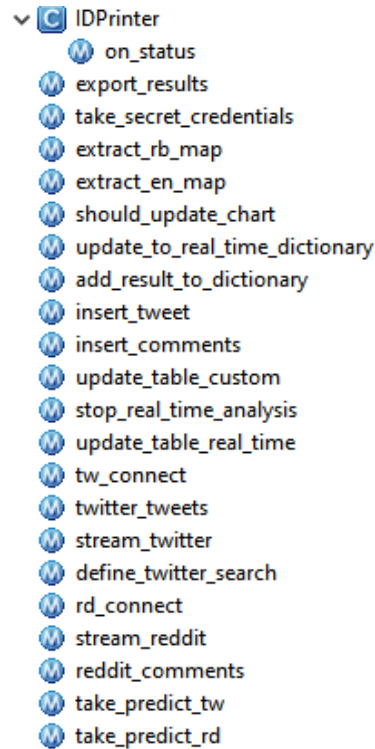


Ilustración 27 - SM-Meter: Métodos en el archivo apiSupport.py

1.1.3. Neural.py

1.1.3.1. Explicación

El archivo neural contiene la lógica de los análisis de los comentarios con los modelos neuronales. Se encarga del control de los modelos.

1.1.3.2. Métodos

Este fichero no contiene ninguna clase.

Contiene los métodos de control de los modelos, tanto de conocer su estado (activado o desactivado) cómo para cargarlos.

También contiene los métodos para preprocesar los comentarios en los idiomas disponibles. También tenemos los métodos para analizar los textos de Reddit o Twitter.

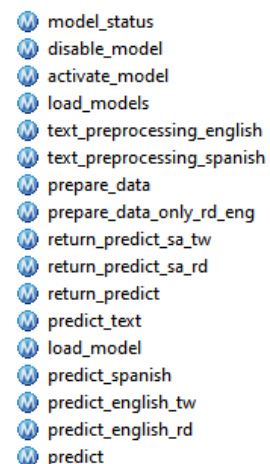


Ilustración 28 - SM-Meter: Métodos en el archivo neural.py

1.2. Problemas encontrados

El gran problema del desarrollo de la aplicación se trata de la necesidad de usar hilos en su ejecución ya que los métodos que son llamados para el streaming no deben ser ejecutados en el hilo principal ya que bloquea toda la interfaz gráfica inutilizando la aplicación por completo, impidiendo incluso para los análisis.

Por ello se ha usado una lógica de multihilos, los métodos de la clase main que llaman a otros métodos de la clase apiSupport se hace creando un nuevo hilo que los ejecute.

Para ello se ha usado la librería *Threading*.

```
lambda: threading.Thread(target=export_results, args=[1]).start()
```

ejemplo de llamada a una

función usando hilos.

2. Flujo de la aplicación

Una parte interesante de la aplicación es ver cómo funciona su análisis general a la hora de realizar un análisis.

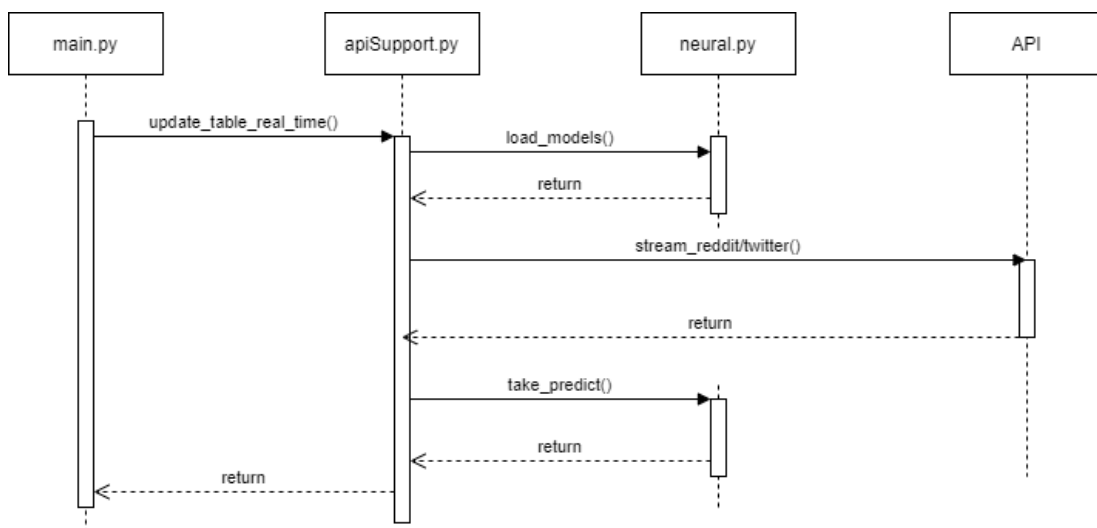


Ilustración 29 - SM-Meter: Diagrama de secuencia

Cuando se realiza una llamada desde main se consulta el archivo apiSupport. El cual analizará la llamada que ha hecho el usuario cargando los modelos estrictamente necesarios. El flujo volverá de nuevo al archivo apiSupport el cual realizará la llamada a la API que corresponda, devolviendo esta los comentarios de los usuarios. Estos comentarios serán enviados de nuevo al archivo neural el cuál los analizará con los modelos cargados previamente.

Al final del análisis (o en mitad de su ejecución) se realizará la llamada para actualizar los gráficos y finalmente se devolverá el resultado final.

3. Capturas finales

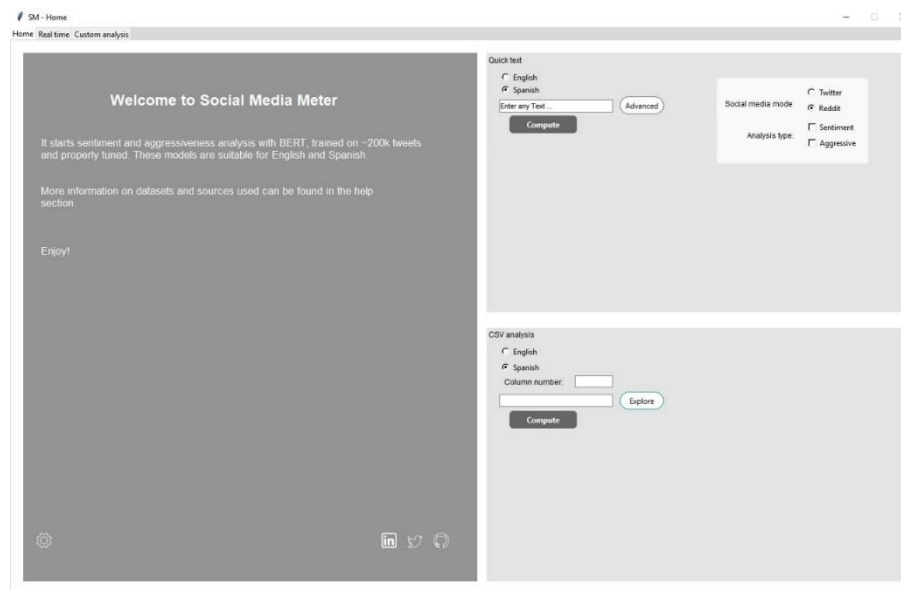


Ilustración 30 - APP: SM-Meter Home

La ventana de home se modificó finalmente añadiendo el input de la columna a analizar en el CSV. El resto quedó (a excepción de algunos colores del diseño) igual a los mockups.

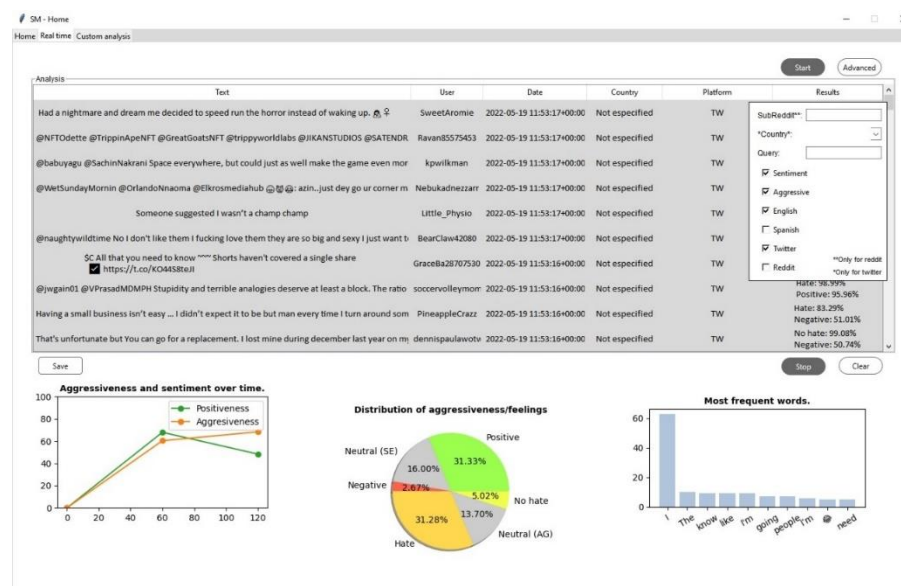


Ilustración 31 - APP: SM-Meter Análisis en tiempo real

La ventana de análisis en tiempo real finalmente contiene los tres gráficos que se detallan en la ilustración superior. Estos gráficos se van actualizando cada 60 segundos y muestra la información en tiempo real según el análisis realizado.

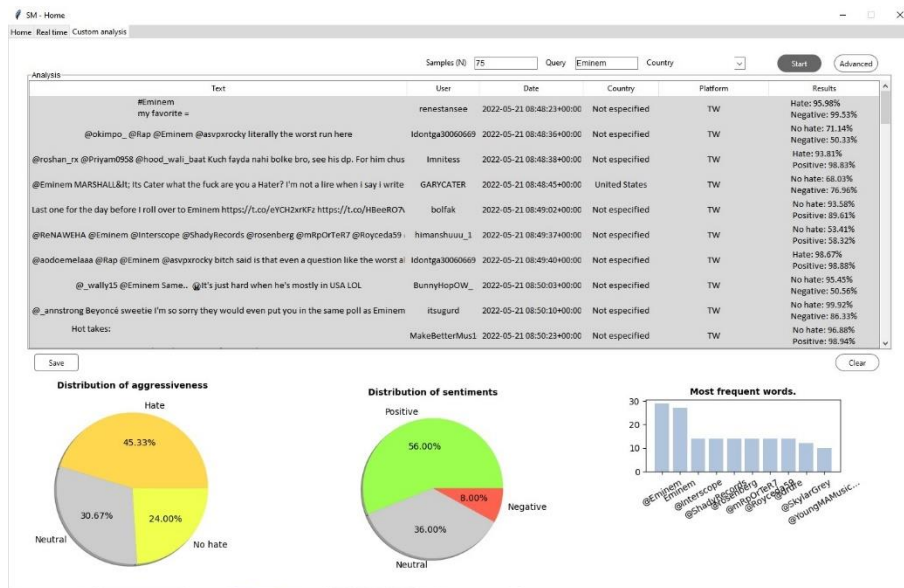


Ilustración 32 - APP: SM-Meter Análisis personalizado

La ventana de análisis personalizado es ligeramente diferente al análisis en tiempo real. Los gráficos se muestran al final del análisis.

La ventana de configuración general muestra los modelos a iniciar o detener. Cuando un modelo está activo el botón se muestra activado.

Se muestran los seis modelos disponibles. Cuando el usuario pasa el puntero por encima de ellos se le informa el estado de estos.

Ilustración 33 - APP: SM-Meter opciones: API

Ilustración 32 - APP: SM-Meter opciones: General

La ventana de API da la posibilidad al usuario de cambiar la configuración con la que se va a interactuar con los servicios de terceros que son utilizados (en este caso Twitter y Reddit). En nuestro caso, por defecto, tenemos APIs gratuitas. Es por ello que este menú es importante si el usuario decide mejorar el uso por defecto de la aplicación. Las credenciales están ocultas hasta que el usuario hace click en el botón con forma de ojo.

Finalmente la parte del menú de información de la app contiene la versión y la fecha al igual que la ilustración de diseño.

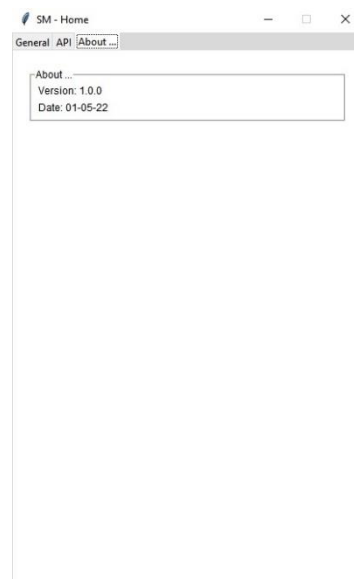


Ilustración 33 - APP: SM-Meter
opciones: About

Los datos son guardados en un formato en csv de la siguiente manera, se guarda el texto junto con su análisis.

text	Sentiment	Aggressive
What about the US ambassador?	Positive: 60.68%	Hate: 99.78%
I'll keep you my dirty little secret, don't tell anyone o	Negative: 89.27%	No hate: 60.39%
@Kim_In_Public @Reddit Going to look for you there	Positive: 99.18%	Hate: 94.83%
Check out S.H.O.N.AðŸŸŸ's video! #TikTok https://	Positive: 98.97%	Hate: 99.02%
On a serious note though, he isnt doing too bad only	Negative: 54.02%	Hate: 97.01%
@AFTVMedia It's that bad i never thought I could see	Negative: 50.34%	Hate: 98.17%
@AngelaBelcamino I was having an MRI- aided biops	Negative: 82.79%	Hate: 95.4%
@SheNastyyAsf Thatâ€™s how you ride a dick	Positive: 98.66%	Hate: 97.76%
4ED5F5B3 :Battle ID need backup!Lvl 150 Proto Baham	Positive: 98.06%	Hate: 99.45%
@HaleeJ At this point but a new wardrobe haha	Positive: 99.11%	No hate: 95.33%
@beomjunology LOVE YOU TOOOO Do u want to see	Positive: 99.06%	No hate: 98.69%
@BryhtDahdze @COTPOfficial You have no real trade	Negative: 68.13%	No hate: 69.74%
@jesskimnft @GreatGoatsNFT @Meme_King1111 @i	Positive: 99.72%	Hate: 99.62%
@bilslimelight girlll my mom takes so many but my a	Negative: 50.85%	No hate: 64.72%
@Lenskart_com Please stop sending messages and re	Negative: 90.73%	Hate: 61.0%
@Oblcubana @TundeEddnut I need 56k#misscampus	Negative: 56.78%	No hate: 99.89%
@DailyMailUK Of course he will. He will be a complet	Positive: 98.88%	No hate: 83.22%
OMG .. @vishy64theking lost the game with #DavidG	Negative: 50.78%	Hate: 99.77%
@ilyfbh @hrrystowel TWITTER HARRIES ARE A WHOL	Positive: 96.19%	Hate: 97.2%
last night i trimmed 100min of BTS footage down to 9	Positive: 84.32%	Hate: 99.38%
Sorry I couldnâ€™t save the world my friend I was toc	Positive: 74.21%	Hate: 98.07%
@LordBarak @LenCarson @violet_octupi @calvinjbui	Positive: 91.14%	Hate: 76.41%
Musk will be president one day, iâ€™m calling it.	Positive: 98.6%	Hate: 95.19%
that means the bear phonecase is coming back on i pl	Positive: 94.72%	Hate: 97.02%
Got up extra early this morning so I can get some laur	Positive: 74.12%	No hate: 91.97%
Forced to endure what I could not forgive	Negative: 54.28%	Hate: 94.69%
@CavalDenis @BitcoinxDaily @Stepnofficial You hav	Positive: 99.74%	Hate: 99.24%
@PokemonGoApp Can you make people move to my	Positive: 72.17%	No hate: 59.34%
@Solimude when you're hyping jay... it's always on r	Positive: 99.82%	No hate: 99.69%
I hate when I smell like weed now everybody wanna	Negative: 50.98%	Hate: 97.09%
I think they will drop a teaser by next week	Positive: 99.43%	Hate: 96.74%
@TomPark1n Projected seats 33 to 72....doesn't seem	Negative: 52.37%	Hate: 99.76%
When did I get so gay? Like tell me why I got a 20 min	Positive: 61.74%	No hate: 99.39%
Iâ€™m so hungry rn	Negative: 58.31%	No hate: 50.57%

Tabla 22 - Exportación de datos en CSV

Conclusiones

1. Conclusiones

Cómo podemos observar se ha desarrollado una aplicación que cumple con los objetivos propuestos. El desarrollo se ha realizado a tiempo y sin mayores problemas.

Tenemos una aplicación altamente adaptable a otros modelos que pueden ser entrenados y modificable a preferencia del usuario por las APIs.

También se contemplan mejoras futuras que se detallarán en el último capítulo.

Enlace al repositorio de GitHub: [ezeperosos/SM-Meter \(github.com\)](https://github.com/ezeperosos/SM-Meter)

Cómo interpretar este capítulo

En este capítulo se expone los resultados de los estudios realizados. Aprovechando los modelos entrenados se va a analizar tweets de dos hechos históricos.

Se van a estudiar los tweets de los hechos que se explican en el apartado Objetivos en la sección **OBJ-03**.

Se procederá de la siguiente manera.

1. Obtención de los datos necesarios
2. Estudio de sentimientos y agresividad
3. Estudio de la naturaleza de los datos

El segundo apartado dependerá de los datos obtenidos y cómo difiere de uno y otro. Se explicará en detalle en cada uno de los apartados.

Los resultados obtenidos se compararán con lo ocurrido finalmente con el tema en cuestión o bien que está ocurriendo en este momento (a día de finalización de este proyecto, la invasión rusa sigue en curso).

Se extraerán las conclusiones una vez realizado todo el estudio.

Datasets

Los datasets que se han buscado para este último capítulo son:

1. Datos hablen de cualquiera de los dos candidatos a las elecciones de EE. UU.
2. Tweets que hagan referencia a alguno de los países implicados en la invasión en territorio ucraniano, ya sea alguno de los países principales, actores principales, menciones de la ONU u opiniones sobre el tema.

Se vuelve a mencionar que los datos han sido obtenidos de un medio gratuito, en este caso Kaggle.

Datasets	Enlace	N.º Tweets utilizados.
US Election 2020 Tweets	US Election 2020 Tweets Kaggle	1.107.929
Ukraine Conflict Twitter Dataset	Ukraine Conflict Twitter Dataset (33.52M tweets) Kaggle	2.160.000

Tabla 23 - Datos utilizados para el análisis

Elecciones EE. UU. 2020

1. Introducción

Las elecciones de estados unidos son un evento importante no solo en norte américa sino para el mundo entero, el cual está pendiente de la jornada electora como de la campaña electoral previa.

Estados unidos es una potencia mundial la cuál afecta en gran medida el panorama internacional y la relación entre países. Un cambio de gobierno en EE.UU. tiene impacto en las relaciones comerciales, internacionales entre países, conflictos bélicos, emergencia climática, etc. Ya que es el tercer país más poblado del mundo según los últimos datos del propio gobierno de estados unidos. (Gobierno EE. UU., s.f.)

Sin ir más lejos las relaciones diplomáticas con Corea del Norte han dado un giro importante después del cambio de gobierno. (Jeremy Diamond, 2022). Con ello también la desnuclearización del país norcoreano. Estas cuestiones las lideraba el gobierno republicano prácticamente cómo portavoz mundial hacia este país y que ahora han quedado en el punto de partida.

El contexto de estas elecciones son con un gobierno repúblicano por parte de Donald Trump desde 2016 que buscaba revalidar su cargo en la presidencia. Por otro lado el candidato demócrata es Joe Biden el cual terminaría ganando estas elecciones para el partido demócrata, partido que no ganaba unas elecciones desde 2012 las cuales venció Barack Obama para seguir en su último mandato en la casa blanca.

Las últimas elecciones de 2016 quedaron empañadas por la famosa polémica de Cambridge Analítica. (Wikipedia, Escándalo Facebook-Cambridge Analytica, s.f.). Según el caso, esta compañía recopiló millones de datos de usuarios de facebook los cuales sirvieron para realizar la campaña de Donald Trump para crear perfiles psicologicos. De esta forma los mensajes llegaban de manera programada y pensada hacia distintos sectores según su clasificación psicologica, el mensaje que contenía también era distinto según el caso. El objetivo no solo era que votasen por el candidato republicano, sino también evitar que se votase por el candidato rival.

2. Por qué es interesante estudiar estas elecciones

Estas elecciones son las siguientes después de la polémica de Cambridge Analytica, por lo que el movimiento en redes se auguraba intenso y vital para cualquiera de las candidaturas. En este caso no solo Facebook está en el punto de mira sino otras redes como la que estudiaremos (Twitter) la cual el candidato republicano era muy activo (De hecho en el momento en el que se está escribiendo este texto Donald Trump sigue baneado de la red social).

De hecho Twitter ya sabía de la importancia que iba a tener su red por lo que presuntamente tomó las acciones necesarias para que los bulos y el arrastre de opinión estuviesen lo más limitado posible.

(Twitter, 2016) Se tomaron medidas tales como:

- Identificación de candidatos

- Ofrecer contexto cuando la información resulte engañosa
- Nueva ventana para los residentes en EE. UU. que contiene información verificada.
- Eliminación de publicidad política
- Asociación con organizaciones para actualizar el estado de las elecciones

Twitter conocía el potencial de su red social además de tener en cuenta la polémica de las elecciones anteriores.

3. Estudio elecciones US 2020

Teniendo en cuenta el contexto explicado anteriormente vamos a analizar los datos que tenemos de twitter.

Los datos se han procesado analizando su positividad y odio utilizando los modelos que hemos entrenado en los capítulos anteriores.

Una forma de obtener información interesante del contenido de los comentarios es realizar la nube de palabras mostrando las palabras más frecuentes de las cuales podemos sacar algunas conclusiones.

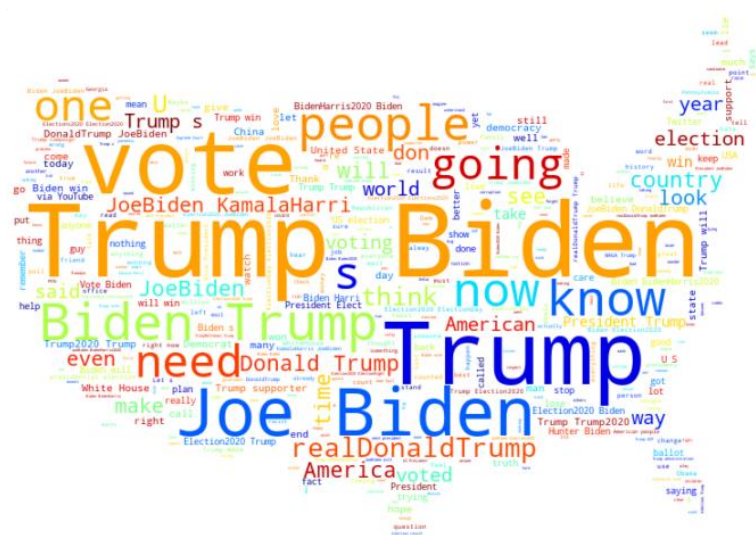


Ilustración 34 - Nube de palabras de las elecciones de EE. UU en Twitter

Como podemos observar, tenemos a los candidatos cómo las palabras más frecuentes en la nube.

En dicha nube podemos destacar el nombre de Kamala Harris, personalidad importante del partido demócrata, ya que se trata de la que terminaría siendo la vicepresidenta de los EE. UU. además de ser la primera mujer en ostentar este cargo. (A parte de ser también la primera persona negra y también la primera persona de ascendencia asiática). Su presencia en la nube de palabras de la ilustración refuerza en cierta manera al candidato demócrata, ya que además de ser un nombre destacado no vemos a ningún participante del bando republicano en dicha imagen.

Si centramos nuestra atención en palabras que muestren algún tipo de emoción, destacan palabras cómo *lie*, *hope*, *love*, *change* y *support*. Estas palabras muestran sentimientos distintos y que han sido frecuente en la opinión de las personas en la red.

China también es nombrado con frecuencia en los comentarios. Es sabido que las relaciones diplomáticas con EE. UU. durante el gobierno republicano se tensaron hasta un punto en que prácticamente quedaron rotas. (Gil, 2021)

La mención constante de este país es un hecho interesante para el contexto en el que se manejaban las elecciones, ya que existieron acusaciones sobre china por parte de Trump y Rusia por parte de Biden sobre intento de manipulación en el proceso electoral. (EuropaPress, 2021) (Arciniegas, 2020). Esto añadiría un punto a favor del candidato demócrata, ya que vemos que es un tema de interés general en la población norteamericana el cuál cómo hemos visto no ha sido tratado de manera correcta por el anterior presidente.

Si filtramos los tweets por candidato (es decir, por un lado las menciones a Biden y otro a Trump) nos queda las siguientes nubes de palabras.



Ilustración 36 - Nube de palabras de las menciones a Trump de EE. UU

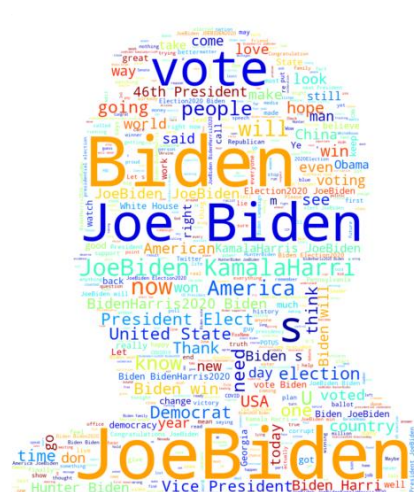


Ilustración 35 - Nube de palabras de las menciones a Biden de EE. UU

Algunos puntos interesantes que hemos obtenido son los siguientes:

Word	Trump	Biden
Freedom	✓	X
Hope	✓	✓
Love	✓	X
Hate	✓	X
Family	✓	✓
Covid	✓	✓
China	✓	✓
Lie	✓	X
Happy	X	✓
Change	X	✓
Corrupt	X	✓
Pennsylvania	X	✓
Proud	X	✓
Georgia	X	✓

Tabla 24 - Palabras más frecuentes en las elecciones

Como se puede ver en la tabla ambos candidatos tienen palabras positivas y otras negativas entre las más comunes. Los temas de Covid y China están presentes en ambos candidatos. Sin embargo hay dos palabras muy importantes que hacen referencia a dos de los estados más importantes en cuanto al marco electoral, que además fueron dos estados que estuvieron en el lado republicano en las pasadas elecciones. En las elecciones presidenciales de los Estados Unidos siempre se ha tenido en cuenta estados que son decisivos y por los cuales los partidos suelen centrar más su campaña. La diferencia de electores de un estado como Pensilvania y Alaska es de 17 a favor del primero.

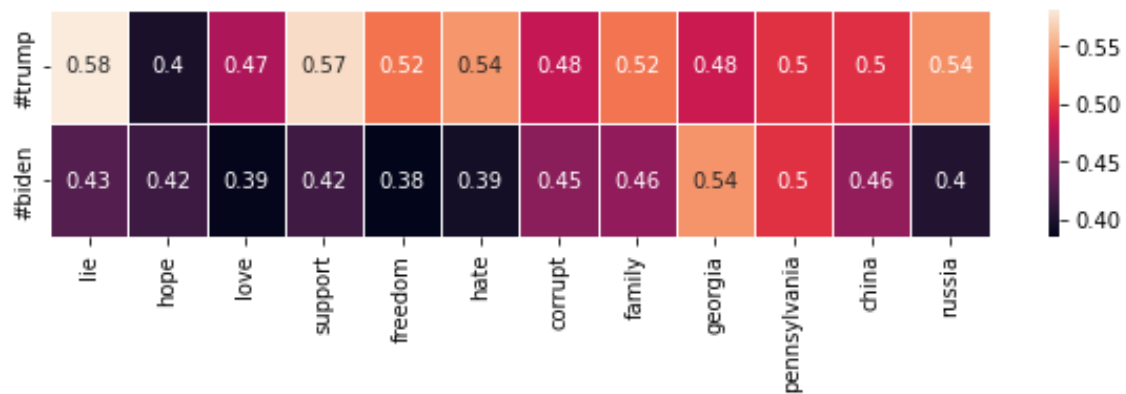


Ilustración 37 - Matriz de correlación de las elecciones de EE. UU.

Aquí podemos observar la relación que tienen cada una de las palabras con los candidatos usando una matriz de correlación, esto es, la relación lineal que tienen dos palabras dadas. Cómo podemos observar tenemos palabras positivas como *love*, *support* y *freedom* (Amor, apoyo y libertad) a favor de Trump, sin embargo tenemos otras palabras negativas como *lie* y *hate* (Mentira y odio) que también son más frecuentes con este candidato. También podemos observar los dos estados antes mencionados y los dos países que más se mencionaron en las elecciones.

Después de un análisis con los modelos predictivos que hemos entrenado podemos ver qué candidato es el más favorecido. Para ello vamos a volver a usar aquellos comentarios donde se nombra a los candidatos.

Una vez que tenemos los datos vamos a ver cuántos tenemos de cada uno de ellos. En la gráfica de distribución de tweets (Ilustración 36) se muestran cuanto se nombran a cada uno de ellos.

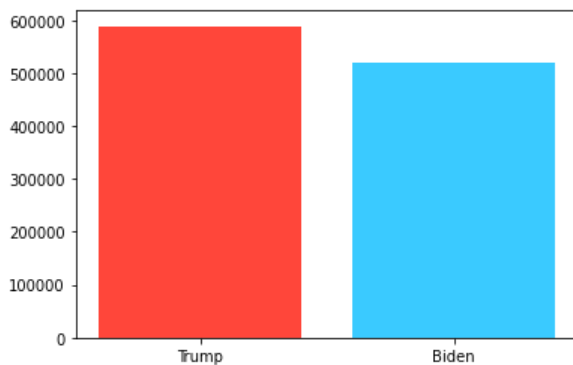


Ilustración 38 - Nº. Tweets de Trump y Biden

Cómo podemos observar los tweets que hacen referencia a Trump son ligeramente mayores a aquellos que lo hacen a Biden.

Por ello el análisis se realizará de la siguiente manera. Se tomarán todos los tweets positivos y se restarán los negativos. Y se hará lo mismo con los comentarios que tienen odio. Una vez hecho esto sacaremos

la gráfica correspondiente para ver los resultados y ver los rangos de cada una de las emociones que hacen referencia a los dos protagonistas del análisis.

Recordamos a este punto del análisis que nuestro algoritmo no es capaz de diferenciar el contexto de una frase, es decir. Si los dos candidatos aparecen mencionados en la misma no podemos saber a quién hace el favor en ese contexto.

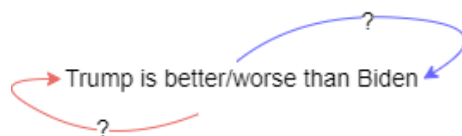


Ilustración 39 - Ejemplo de comentario hacia uno de los dos candidatos

Aunque los comentarios se han obtenido según las menciones a los candidatos no podemos separar totalmente los comentarios de ambos, por lo que debemos tener en cuenta que tendremos algo de *ruido* en la predicción final.

Después de realizar el análisis obtenemos los siguientes resultados:

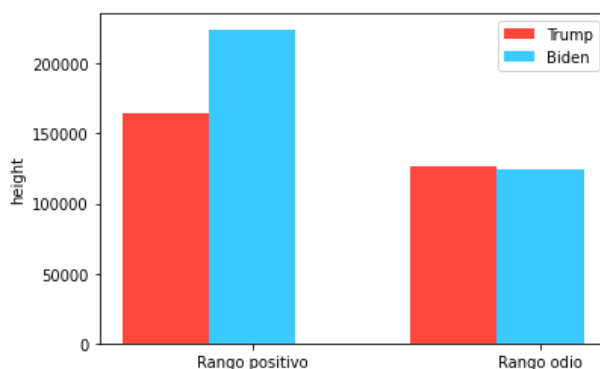


Ilustración 40 - Rango de positividad y agresividad de los comentarios recibidos por los candidatos

Cómo podemos apreciar el candidato republicano tiene menos positividad en los comentarios recibidos que el candidato demócrata, por otra parte, en los comentarios de odio hay una cierta igualdad. Esto se podrá observar mejor en los mapas de calor que podemos generar con el código de cada estado en los comentarios que hemos recogido en twitter.

Estos comentarios con el código de estado son inferiores, ya que no todos tiene la geolocalización activada.

De hecho podemos observar cómo el número de comentarios con código de estado de parte del candidato demócrata son mayores. Para paliar esto vamos a sacar las relaciones entre los comentarios positivos/sin odio y los comentarios negativos/odio. La diferencia de la relación entre los candidatos en cada estado será la que determine el resultado.

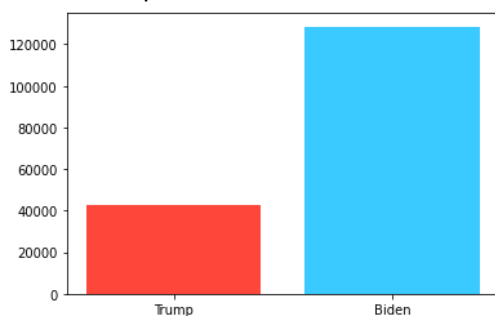


Ilustración 41 - Nº. Tweets con geolocalización activada por candidato

Ahora observemos el mapa de calor de los resultados del análisis usando el campo del código de estado primero con los comentarios que expresan positividad y negatividad.

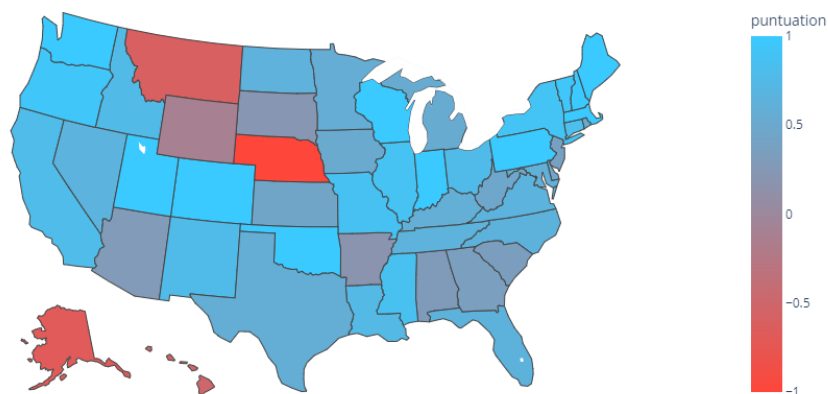


Ilustración 42 - Mapa de resultados tras el análisis de sentimiento

Como podemos observar el candidato demócrata tiene muchos más comentarios positivos que el candidato republicano. Sin embargo a estos resultados también le tenemos que añadir los comentarios de odio.

Por otra parte los comentarios de odio se distribuyen de esta manera.

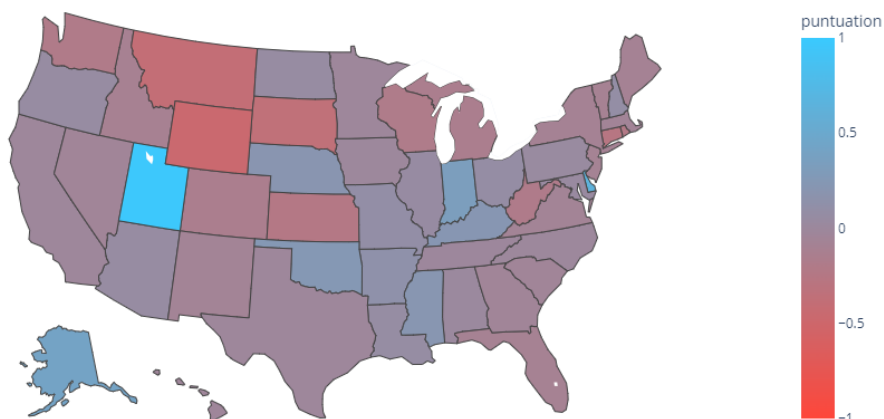


Ilustración 43 - Mapa de resultados tras el análisis de agresividad

Como podemos observar los comentarios de odio son más frecuentes al bando republicano. Sin embargo, comentarios o discursos de este tipo han sido también un arma utilizada por este

partido frecuentemente (de ahí a que su candidato esté baneado de la red). Si fusionamos ambos mapas nos queda el siguiente resultado.

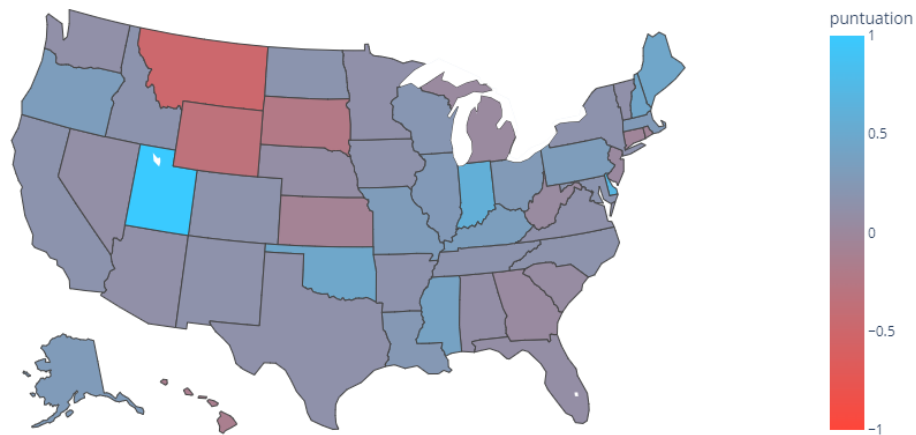


Ilustración 44 - Mapa de resultados tras el análisis completo

Cómo podemos observar no hay demasiados estados que se decanten de forma clara por uno u otro, de hecho esto fue una realidad en las elecciones, ya que la mayoría de los estados se ganaron de forma muy justa. (CNN, 2020) Podemos observar cómo hay estados que caen del bando republicano de forma clara como Montana, Wyoming o Dakota en los cuales si hubo una mayor diferencia, también en el bando demócrata con Illinois, Oregon, Nuevo Hampshire. Hay estados como Utah que a pesar de darle una clara victoria al partido demócrata fue finalmente lo contrario.

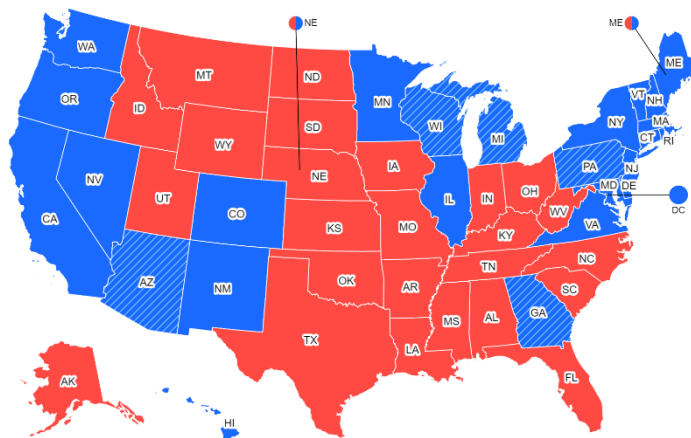


Ilustración 45 - Resultado final de las elecciones de EE. UU. de 2020

El mapa final de las elecciones presidenciales quedó de la siguiente manera. El resultado puede ser engañoso porque da una sensación de superioridad en todos los estados de cada candidato, sin embargo, como se ha puntualizado, esto no fue así y se tuvo una victoria ajustada. De hecho el número de votos final fue de 81,284,666 de votos para Biden y 74,224,319 para Trump, haciendo la victoria del demócrata por un 51.3%

4. Conclusiones

Como hemos podido observar, el análisis de redes sociales nos puede dar datos muy potentes sobre temas de actualidad. En este caso en las elecciones de EE. UU. Los datos coinciden con el resultado final. Biden terminó ganando las elecciones y dando la vuelta a estados importantes como Pensilvania y Georgia las cuales se hace mención en los Tweets recogidos.

Además con los datos mostrados podemos ver como se ha predicho la igualdad resultante en cuanto a votos y en estados. También coinciden los resultados con los discursos polémicos y duros de Donald Trump, lo cual se deja ver en los mapas de calor de odio. Ilustración 43 - Mapa de resultados tras el análisis de agresividad. Tal y como hemos podido observar no se deja atrás los conflictos con otros países los cuales están presentes en los comentarios de los usuarios y nos deja ver que es un tema importante para la población, lo cual nos deja pistas de qué discurso o política podría tener más éxito cómo candidato.

Este análisis nos demuestra que las encuestas electorales no son el único medio para obtener información de la población. En redes sociales los usuarios son más atrevidos por el falso anonimato y muestran sus opiniones sin tapujos, dando como resultado que estos análisis tengan un gran éxito. Además, los rangos de edades que usan las redes son cada vez mayor. Esto se debe a que cada vez hay más presencia de edades distintas en las redes debido al auge de las TICs en edades más avanzadas y la lucha por minimizar la brecha digital. (INE, 2021)

Finalmente, no es de extrañar las políticas que se están tomando cuando hechos manipulables son más controlados por parte de las instituciones o empresas que llevan estas redes.

Esto también nos hace ver también el peligro que puede tener una noticia falsa, ya que esta puede moverse fácilmente en este campo y llegar a “contaminar” la opinión pública, siendo esto un boca a boca mucho más difícil de parar, ya que vivimos en una corriente constante de datos que nos hace difícil diferenciar quién está diciendo la verdad o quién no.

Las redes han terminado siendo también una herramienta política, de ahí a que los partidos políticos estén más presentes cada vez en internet y además crezcan a través de este medio.

Invasión Rusa

1. Introducción

El 24 de febrero de 2022 Rusia inicia una operación militar en Ucrania excusados en desmilitarizar el país y desnazificarlo, la guerra explota en fronteras cercanas a Dombás, Crimea y Rusia. Es el inicio de la invasión Rusa en territorio Ucraniano que sigue vigente tres meses después de su inicio.

Este evento tiene unos antecedentes que llegan hasta los tiempos de la unión soviética, pero nos basta con saber que parte de él nace por los conflictos en las regiones del Dombás y Crimea, también por la idea de la anexión de Ucrania a la OTAN y por la influencia Rusa sobre el territorio ucraniano.

Durante todo el conflicto desde su inicio no se ha parado de comentar en redes sociales, mostrando apoyo a uno u otro bando, comentando el día a día de la guerra o también intentado informar o desinformar sobre los progresos de ambos. La cantidad de mensajes y opiniones que circulan en redes sociales es abrumadora sobre todo en los días de comienzo del conflicto.

Por ello vamos a analizar alrededor de 2.2M de tweets que van desde el día 24 de febrero a 13 de Abril de 2022. Se sacarán los movimientos de las opiniones en redes sociales cómo a su vez la reacción a los eventos que van ocurriendo día a día. Podremos observar cómo el movimiento de la guerra y sus eventos también arrastran el movimiento de la opinión.

2. Matices del estudio

Es importante destacar algunas condiciones a las que está sujeto este estudio.

Desde la primera semana de marzo twitter ha sido bloqueado (entre otras redes sociales) en Rusia. Esto quiere decir que los usuarios rusos no pueden exponer su opinión de forma habitual ya que la propia Rusia ha bloqueado el acceso a las redes. Esto afectará al estudio ya que no tendremos opiniones de personas que residen en este territorio (Aunque vivan en este país no podemos afirmar que por ello estén a favor en este evento). (Los angeles times, 2022)

También tenemos que tener en cuenta que hemos tomado 80k tweets por día de forma aleatoria y que han sido escritos en inglés, por lo que no tenemos los datos absolutos sino una generalización desde el habla inglesa (ya que nuestros modelos soportan este lenguaje).

Por último, la mayoría de opiniones provienen de países occidentales. Por lo que las opiniones de países de asia oriental que están mas influenciados por Rusia se nos escapan en los datos.

3. Estudio invasión Rusa

Con lo explicado anteriormente podemos exponer el análisis realizado.

Si formamos una nube de palabras con todos los comentarios entre las fechas indicadas anteriormente, nos queda la siguiente ilustración. Recordemos que en esta nube aparecen los términos más frecuentes.



Ilustración 46- Nube de palabras de los datos de la invasión rusa

Como podemos observar encontramos muchos términos interesantes como los nombres de los países implicados además de los gentilicios de dichos países, las menciones de la propia guerra, ciudades (Cómo Kyiv y Mariúpol) y otro tipo de palabras que muestran o bien el rechazo a la guerra o el rechazo a rusia o sus dirigentes.

Se pueden encontrar las siguientes palabras, *family, children, innocent, refugee, freedom o killed* que nos dan una idea del contexto de los comentarios que estamos tratando. Podemos encontrar también palabras cómo *army, destroyed o aggression* que son palabras propias de un evento como este.

Con solo observar esta nube de palabras podemos darnos cuenta de lo que estamos tratando sin siquiera conocer el contexto de este.

Un buen punto que podemos observar es de donde se están enviando estos mensajes, es por ello que aquellos con la localización activada nos permite saber este dato. Para ello construimos la siguiente gráfica.

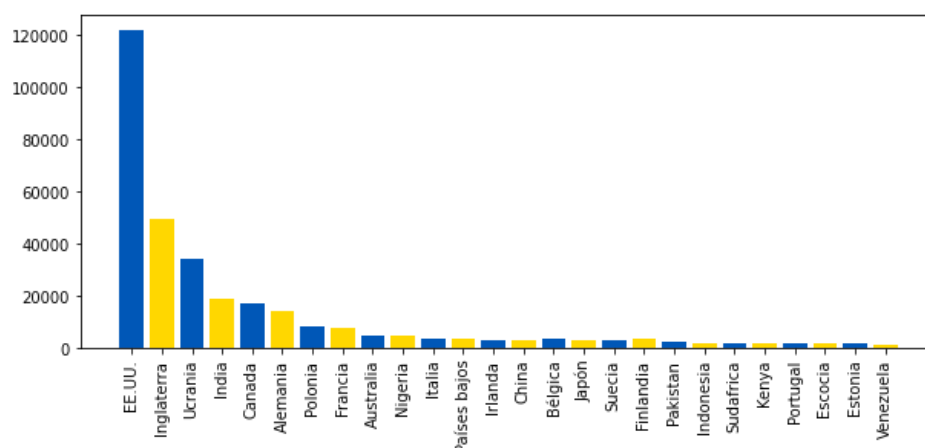


Ilustración 47 - Distribución de los comentarios por país.

Como podemos observar la gran mayoría de comentarios son de EE. UU. E Inglaterra. Algo que no es de extrañar si estamos procesando comentarios en inglés. El tercer puesto recae en Ucrania, probablemente por todos los medios que cubren las noticias desde el país y también por un intento de que la información u opiniones de los ciudadanos lleguen a la mayor cantidad de gente posible. Podemos ver que Rusia no aparece entre los países que más comentarios envían, esto puede deberse a que aparte de que ha sido censurada en el país tal y como indicábamos en el punto anterior, el número de personas de la población rusa que ha usado esta red social no es muy grande. (14% aproximadamente (Roa, 2022)).

Ahora bien, cómo hemos visto en la lista de palabras, tenemos distintas palabras que se usan para expresar una idea distinta. Por ello vamos a clasificarlas de la siguiente manera.

Movimiento	Palabras	Movimiento	Palabras
Protesta guerra	# peace	Apoyo Ruso	#istandwithrussia
	#stopputin	Apoyo Ucraniano	#stopwarinukraine
	#stopwar		#freeukraine
	#stopthewar		#standwithukriane
	#stopputinswar		#saveukraine
	#russiagohome		#standtogether
Anti Putin	#putinwarcriminal		#weareallukrainians
	#putinisawarcriminal		#armukrainenow
	#putinswar		
	#putinhitler		
	#fckputin		

Tabla 25 - Palabras agrupadas por movimiento social

La evolución de estos movimientos sociales puede verse en la siguiente gráfica.



Ilustración 48 - Evolución de los movimientos sociales con el tiempo

En general se puede observar los mensajes contra la guerra siempre han sido muy superiores al resto. Los mensajes contra Putin y el apoyo a Ucrania han sido muy parejos en el tiempo en ocasiones. Por último, el apoyo ruso no ha sido popular en los comentarios, tampoco se han encontrado otras palabras o hashtags a favor de estos. Una vez que presentemos el resto de las gráficas veremos la evolución de estos movimientos a través del tiempo, observando los hechos que han ocurrido y cómo han evolucionado.

Como podemos ver a continuación, estas palabras están relacionadas entre sí, si observamos la matriz de correlación.

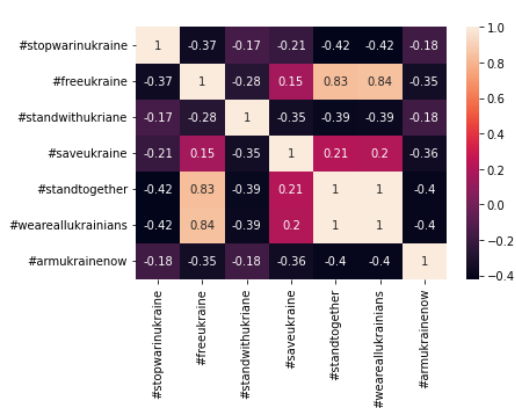


Ilustración 49 - Matriz de correlación de apoyo a Ucrania

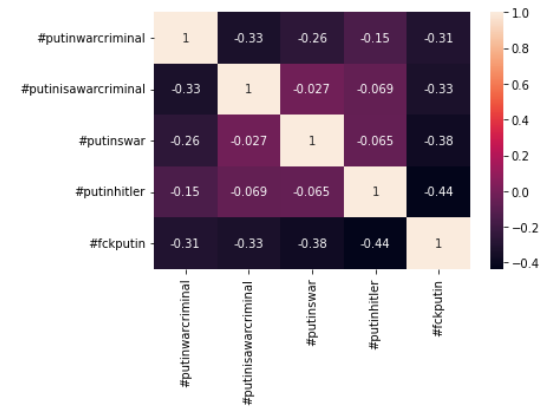


Ilustración 50 - Matriz de correlación contra Putin

Podemos observar que los grupos de palabras suelen venir acompañadas de otras con el mismo mensaje, esto es común cuando se intenta transmitir una idea y reforzarla con términos similares en el mismo mensaje.

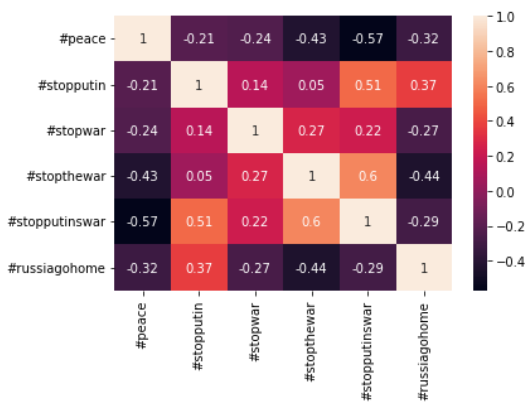


Ilustración 51 - Matriz de correlación a favor de para la guerra

Tenemos otras palabras claves que son las siguientes.

Palabra	Palabra
#anonymous	#mariupol
#biden	#nato
#bucha	#Slavaukraini
#kharkiv	#chernobyl
#kyiv	#belarus

Tabla 26 - Palabras seleccionadas para la gráfica de eventos

Como podemos observar tenemos palabras que se mencionan eventos que ocurrieron durante estos días, desde ciudades atacadas, instituciones importantes y otros países de interés.

¿Cómo se han sacado estas palabras? Para ello se han tomado las palabras más frecuentes durante estos días y se han seleccionado las más importantes. Se han dejado fuera las palabras que se mostraban en la gráfica superior ya que no destacan hechos sino sentimientos de los usuarios.

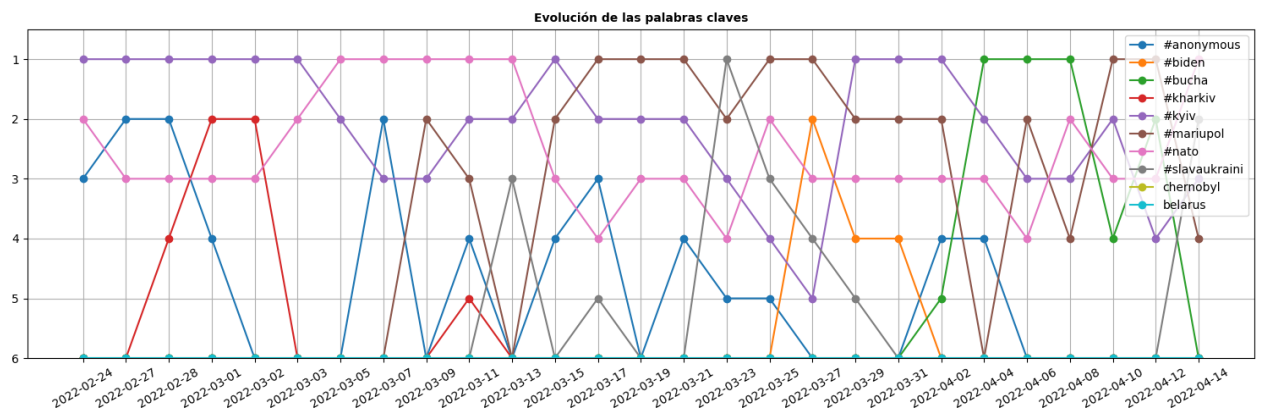


Ilustración 52 - Evolución de las palabras claves en el tiempo

Cómo podemos observar en la gráfica, la evolución de las menciones de las palabras claves no son constantes y tienen subidas y bajadas dependiendo de lo que está ocurriendo en ese momento, tal y como veremos más adelante.

Otras palabras claves se han dejado en otra gráfica aparte, estas hacen referencia a instituciones o eventos que no se mantienen en el tiempo pero que tienen picos de menciones que son interesante estudiar.



Ilustración 53 - Evolución de las palabras de eventos en el tiempo

Cronológicamente podemos observar lo siguiente (Wikipedia, 2022) (CNN, 2022):

24 de febrero al 3 de marzo

- Hechos:
 - Entrada de tropas rusas a territorio ucraniano.
 - Ataques a distintas ciudades por toda ucrania.
 - Kiev y Járkov son bombardeadas.
 - Se toma la planta nuclear de Chernóbil.
 - Se bombardean aeropuertos y otros sitios estratégicos.
 - Desde Europa se anuncia la expulsión de los bancos rusos del sistema SWIFT.
 - Se pide la anexión de Ucrania a la unión europea.
 - Se toma la ciudad de Jerson.
 - Se menciona la posibilidad de una hipotética tercera guerra mundial.
 - Empiezan las migraciones masivas de personas ucranianas a otros países.
- Análisis:

Podemos ver cómo las palabras NATO (OTAN), Kyiv y Kharkiv (Járkov) ocupan las primeras posiciones. Esto está unido a cómo se dispara los mensajes de parar la guerra en redes sociales. También la palabra Europa se dispara en esta primera semana.

También se nombra a Anonymous después de una filtración masiva de datos de personalidades rusas.

En el análisis de sentimiento tenemos las menciones a Járkov que ha sido una de las ciudades afectadas en la primera semana de guerra. Cómo podemos ver el odio de los mensajes en la primera semana que contienen dicha palabra son prácticamente la mitad de los mensajes enviados que la contienen. Por otro lado los mensajes son mayoritariamente negativos en su contenido, lo cual es esperado por la situación que se está dando. No es lo habitual si nos fijamos en los gráficos que encontraremos más adelante, aunque algo natural en el contexto que estamos tratando.

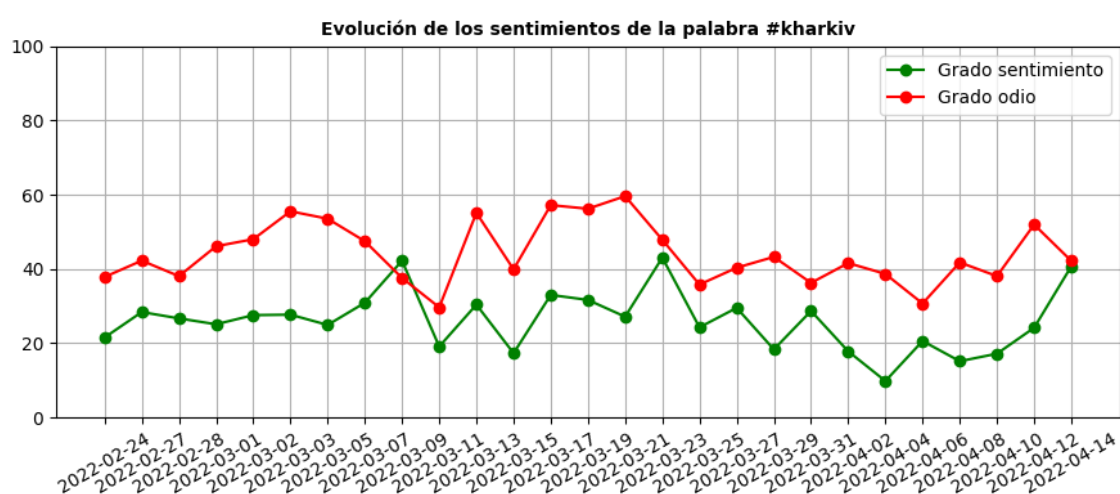


Ilustración 54 - Evolución del grado de positividad y odio de la palabra Kharkiv

4 de marzo al 11 de marzo:

- Hechos:
 - Primeros corredores humanitarios que terminan en una misión fallida.
 - Captura de bucha por parte del ejército ruso.
 - Tercera ronda de negociaciones entre ambos países.
 - Los combates se trasladan al norte de Kiev.
- Análisis:

La OTAN sigue siendo una mención común en los mensajes y crecen las menciones a las ciudades de Kiev y Mariúpol, ciudad donde falló el corredor humanitario. Los comentarios sobre Járkov bajan a favor de Mariúpol. Se pide que se proteja el aeropuerto de Kiev que también fue atacado durante esta semana.

Al final de esta semana empezamos a ver cómo suben el apoyo al pueblo ucraniano y los comentarios contra Putin, los mensajes contra la guerra experimentan una pequeña bajada, probablemente por el enfriamiento inicial del tema en redes sociales.

En el análisis de sentimiento tenemos las menciones a Mariúpol que empieza a ser atacada de forma constante durante estos días. Como podemos ver, los mensajes con contenido de odio siguen predominando sobre aquellos que contienen mensajes positivos, aunque estos crecen, probablemente por los mensajes de apoyo a la ciudad tras los bombardeos continuos. Tras situar el foco en esta ciudad también notamos el cambio de tendencia en las menciones a Jerson, pues los comentarios de odio bajan y suben aquellos positivos, probablemente en un intento de retomar la estabilidad habitual hasta que las noticias se centren en esta ciudad.

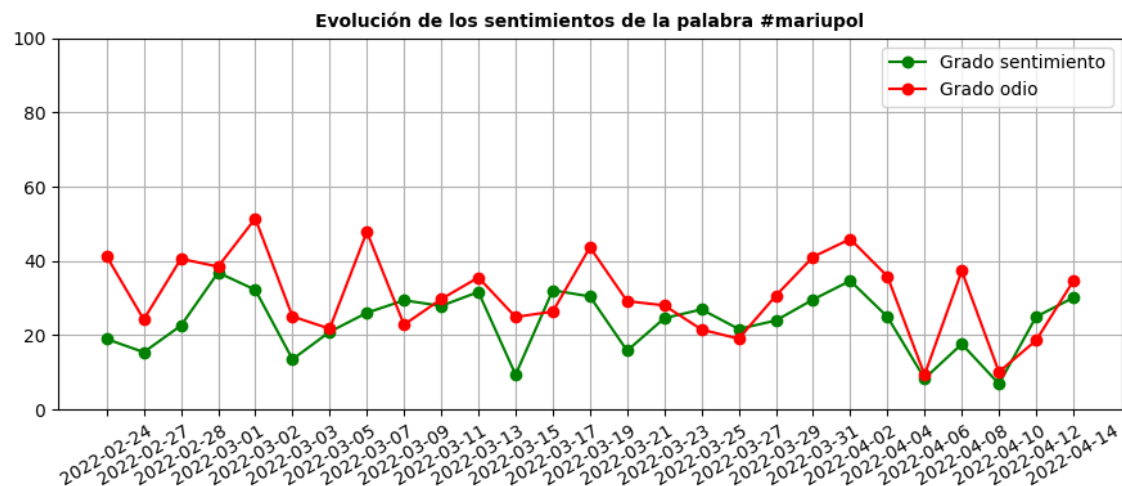


Ilustración 55 - Evolución del grado de positividad y odio de la palabra Mariúpol

12 de marzo al 19 de marzo:

- Hechos:
 - Protestas en la ciudad de Mariúpol por la liberación de su alcalde, Ivan Fedorov.
 - Se registran las primeras amenazas a civiles, cómo también víctimas, por parte de los militares rusos.
 - Se registran muertes de niños ucranianos en la ciudad de Mariúpol, además de un bombardeo a un teatro donde había gran cantidad de estos.
- Análisis:

Cómo podemos observar Mariúpol se sitúa en la primera posición a partir de esta semana debido a los terribles hechos y la gente responde comentando sobre estos bombardeos. También se dispara el apoyo a Ucrania y el rechazo a Putin. Además de Mariúpol, Bielorrusia también es un tema comentado por la posición cercana a Rusia durante el conflicto. Se dispara el hashtag en apoyo a Ivan Fedorov.

En el análisis de sentimiento tenemos que las menciones a Jerson siguen en la misma línea que durante las semanas anteriores. En la ciudad de Mariúpol experimentamos un pico de negatividad, probablemente debido a los bombardeos al teatro y el fallecimiento de los cientos de niños a causa de los bombardeos rusos. Por otra parte, respecto al análisis de la ciudad de Kiev, no registramos cambios importantes. Esto probablemente se deba a que, si bien es cierto que ha estado bajo ataque constante, es la capital del país por lo que no se espera por entonces que el ejército ruso sea capaz de tomarla. Otro dato interesante es que esta ciudad es el origen del pueblo ruso por lo que estos no verían con buenos ojos destruir

una ciudad tan significativa para ellos, por lo que aunque existen bombardeos a esta ciudad, no son tan intensos como en otras del mismo país.

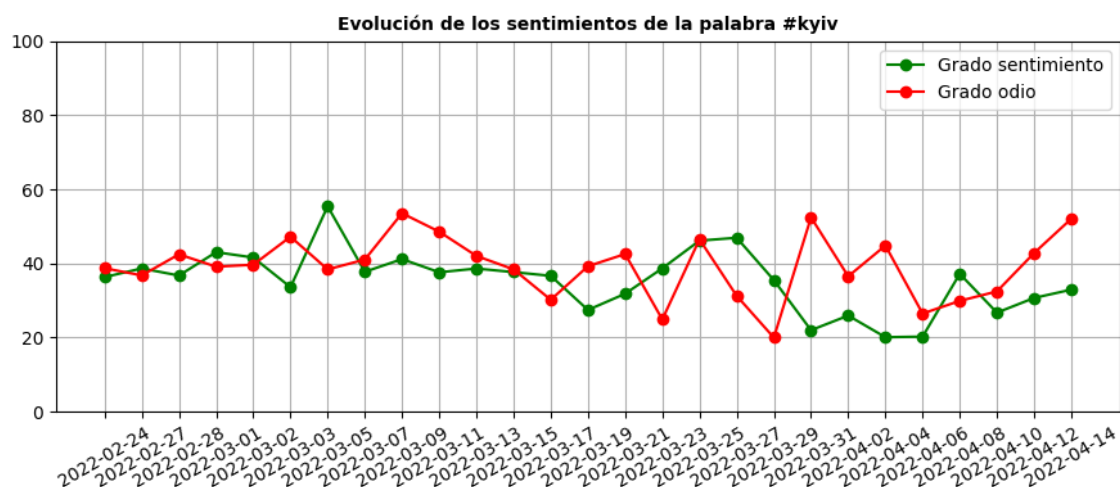


Ilustración 56 - Evolución del grado de positividad y odio de la palabra kiev

20 de marzo al 27 de marzo:

- Hechos:
 - Contraofensiva en Kiev y Járkov por parte del ejercito ucraniano.
 - Se registran suburbios en Bucha.
 - Continúan los bombardeos en Mariúpol.
 - Bajada del número de noticias y su constancia en los medios como venía siendo habitual.
 - El ejército ruso no avanza según sus planes.
- Análisis:

Se vuelve a registrar Mariúpol como la ciudad que más se menciona en las redes además de la OTAN. Los usuarios siguen pidiendo acciones a la OTAN y se sigue mencionando a Anonymous por las amenazas que siguen surgiendo en redes sociales hacia Rusia. Durante esta semana también se registra un aumento de menciones sobre una carta mencionada por Canada, luego dicho país muestra su apoyo a Ucrania.

Esta semana además es la última de mayor movimiento en redes sociales en apoyo a uno de los dos bandos, como vemos la atención cae después de un mes.

En el análisis de sentimiento tenemos que las menciones a Kiev tienen una pequeña curva positiva por las noticias del frente, mientras que en Mariúpol ocurre lo contrario tras los constantes bombardeos que se siguen sucediendo.

28 de marzo al 4 de abril:

- Hechos:
 - Negociaciones en Turquía que terminan fracasando.
 - El ejército ruso se estanca en Kiev el cuál se termina retirando hasta Chernóbil.
 - Ocurre la masacre de Bucha desde donde llegan imágenes de cadáveres civiles y fosas comunes en la ciudad ocupada por el ejército ruso.
- Análisis:

Mariúpol deja de ser mencionado constantemente y deja paso a Kiev, ciudad de la que el ejército ruso se está retirando. Bucha sube de puestos situándose la primera al final de la semana. La palabra genocidio también escala hasta la primera posición.

Se produce una bajada de los movimientos de apoyo que se comienzan a estabilizar por la mitad de sus mayores picos.

En el análisis de sentimiento tenemos que las menciones a Kiev tienen una pequeña curva de bajada por parte de los comentarios de odio, sin embargo los comentarios positivos no terminan de despegar. Los comentarios sobre Bucha tocan su segundo mínimo (el primero fue tras la toma de la ciudad) en los comentarios positivos.

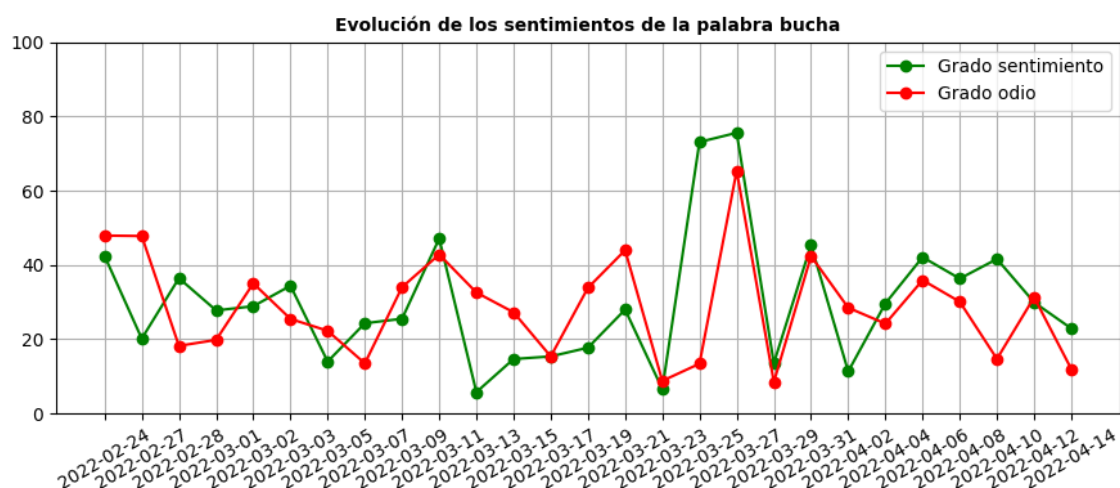


Ilustración 57- Evolución del grado de positividad y odio de la palabra bucha

5 de abril al 14 de abril:

- Hechos:
 - Se siguen registrando víctimas civiles.
 - Sigue el shock tras lo ocurrido en Bucha.
 - Se acusa a Putin de crímenes de guerra.
 - Se pide ayuda militar a Europa.
 - Se hunde el buque Moskva.

En esta semana se siguen registrando víctimas civiles además del reciente shock por lo sucedido en Bucha. Se acusa a Putin de crímenes de guerra y se pide ayuda militar a Europa. Al final de semana se hunde el buque Moskva.

- Análisis:

Bucha sigue siendo el tema principal durante la semana por el genocidio perpetrado por el ejército ruso. Mariúpol y Kiev siguen siendo menciones constantes con un pico de apoyo a Ucrania. Las menciones al buque Moskva cierra la semana. Los movimientos se mantienen como la semana anterior, mostrando el claro enfriamiento en la red.

En el análisis de sentimiento tenemos que la positividad mencionando a Bucha sigue en caída libre. Por otra parte el análisis de la palabra nato nos deja comentarios positivos en general, superando incluso a los comentarios con odio.

Para finalizar, nos damos cuenta de que Jarków vuelve a subir sus comentarios positivos a lo habitual tras terminar el enfoque de las noticias hacia esta ciudad. Las tendencias de los rangos en cada palabra parecen no moverse demasiado tras una semana sin noticias importantes en el frente.

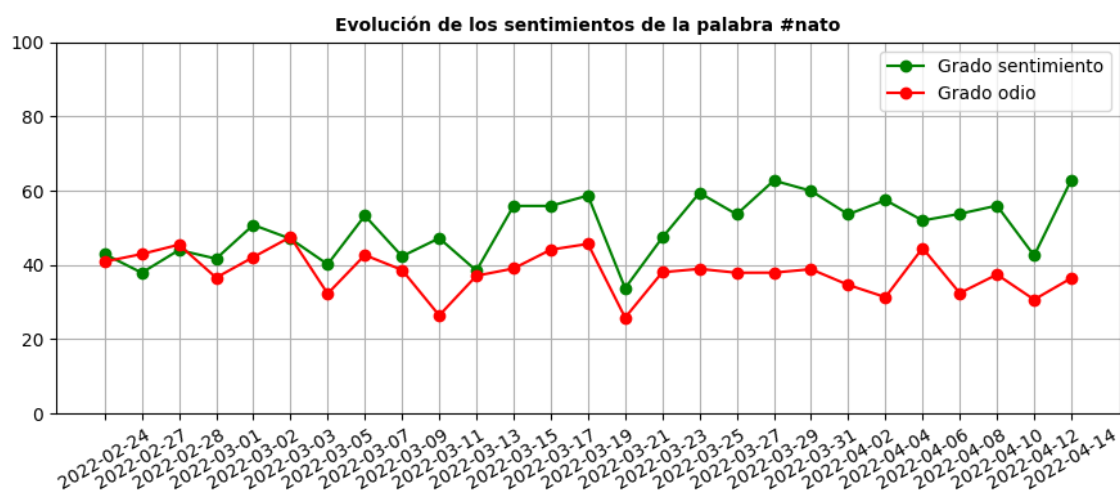


Ilustración 58- Evolución del grado de positividad y odio de la palabra nato

4. Conclusiones

Como hemos podido observar, el análisis de redes sociales nos puede orientar sobre la percepción de un tema de actualidad desde el punto de vista de la población general. En este caso el tema a tratar ha sido un tema complicado, duro e incluso difícil de creer que ocurra en la actualidad. Sin embargo este análisis nos indica que el ser humano se mueve por impulsos en redes sociales, es decir, queremos compartir nuestra opinión sobre aquello que vemos en las noticias, más si es un tema que nos genera un sentimiento de injusticia como este caso (ya que la población que está sufriendo las consecuencias poco tiene que ver con los intereses políticos). Como hemos observado en el análisis, los temas de conversación fluyen a medida que la noticia torna hacia otro punto, es por que la opinión ha sido afectada a medida que se iban conociendo hechos.

Este tipo de análisis también puede ser utilizado por los grupos políticos antes de tomar alguna decisión polémica y así ver la tendencia sobre lo que rodea esta posible decisión.

Por último, también podemos usar este tipo de herramientas para otro tipo de cosas, por ejemplo, en el caso de un inversor que está a punto de realizar una operación muy importante en una organización la cuál está en el punto de mira en las noticias del día a día. Podemos observar así si ese producto está siendo apoyado o no, o si por otro lado, nos conviene realizar cierta operación.

Consideraciones finales

1. Tiempo empleado en la realización del proyecto.

	Título tarea	D. Optimista - Más probable - Pesimista	Tiempo empleado
Sprint 1	Lluvia de ideas	180 – 360 - 480	330
	Selección de ideas	30 – 120 – 240	55
	Reunión Product Owner	30 – 60 – 120	58
	Alcance	40 – 120 – 200	25
	Justificación	30 – 60 – 80	22
	Interesados	30 – 60 – 80	43
	Objetivos	25 – 120 – 160	43
	Requisitos de alto nivel	60 – 120 – 180	100
	Límites	30 – 60 – 80	52
	Planificación temporal	240 – 480 – 600	150
	Planificación de costos	60 – 120 – 200	56
Total Sprint		755 – 1680 – 2420	934
Sprint 2	Recopilación de datos de Twitter	120 – 300 – 400	162
	Recopilación datos de Reddit	120 – 300 – 400	122
	Reunión Product Owner	150 – 240 – 300	85
	Procesado y análisis de datos recolectados	240 – 360 – 500	351
	Construcción del modelo de análisis y sentimiento	480 – 840 – 920	1049
	Entrenamiento y optimización	300 – 540 - 700	1221
	Análisis de los datos recolectados	240 – 360 - 480	300
Total Sprint		2490 - 4380 - 5450	3290
Sprint 3	Diseño de la aplicación	300 – 360 - 400	300
	Configuración API Twitter	260 – 420 - 480	288
	Configuración Reddit	300 - 420 - 480	258
	Programación de app	900 – 1200 – 1500	1828
	Reunión Product Owner	120 – 180 – 200	60
	Pruebas	360 – 540 - 620	250
Total Sprint		2240 – 3120 - 3680	2976
Sprint 4	Recopilación de datos de Twitter del estudio	300 – 480 – 520	147

Título tarea		D. Optimista - Más probable - Pesimista	Tiempo empleado
	Recopilación datos de Reddit del estudio	300 – 480 – 520	*
	Procesamiento de datos	630 – 960 – 980	4300**
	Análisis de los resultados	340 – 540 – 600	302
	Informes y conclusiones	500 – 960 – 1000	621
	Reunión Product Owner	100 – 120 - 180	40
Total Sprint		2170 – 3540 - 3800	5410
Sprint 5	Retrospectiva y optimización general	480 1200 – 1400	420
	Memoria TFG	720 – 1440 – 1600	1450
	Reunión Product Owner	80 – 120 - 180	25
Total Sprint		1280 – 2760 - 3180	1895
Total estimación en horas (Redondeado)		149h - 258h - 309h	241h***

Tabla 27 - Estimación de tareas

Consideraciones:

* No se han recopilado datos de Reddit finalmente para los estudios de investigación.

** El procesamiento de datos ha tomado mucho más tiempo del indicado. Sólo se ha contabilizado los minutos en activo e interactuando con el proyecto durante esta tarea. Este tipo de tareas se realizan de forma automatizada.

*** No han sido añadidas las horas de inactividad porque no han sido recopiladas.

Finalmente respecto a la duración más probable tenemos una desviación de 17 horas positivas, lo cual nuestros tiempos totales han sido bien calculados ya que la desviación ha sido pequeña.

2. Conclusiones generales.

Tras finalizar este proyecto podemos mirar todo el trabajo realizado y concluir que el aprendizaje ha sido muchísimo mayor del esperado. En una titulación donde este campo no se trata en profundidad es de agradecer un trabajo como este. Partiendo de una base prácticamente inexistente en este campo el trabajo de investigación ha sido muy profundo y estudiando muchísimo contenido sobre el tema.

El recorrido desde PLN, programación en Python usando interfaces gráficas, uso de librerías de tratamiento de datos y herramientas para realizar los estudios, deja una satisfacción personal enorme.

Cómo última reflexión personal. Finalizo este proyecto tras un trabajo duro pero en el que me lo he pasado genial aprendiendo de ello, y es que al poner punto final a este documento, me llevo muchísimo conocimiento del que me siento orgulloso.

3. Trabajo futuro.

En un proyecto como este siempre hay margen de mejora. Las consideraciones a mejorar se detallan a continuación. Respecto a los modelos podemos mejorar su desempeño utilizando datasets más profesionales y de pago, aquellos que nos pueden dar un salto de calidad importante y que el corpus de este sea mucho más profundo. Podemos usar un equipo más potente para poder hacer redes más complejas y que el tiempo de entrenamiento no se vaya de los tiempos establecidos. También se podría realizar un solo modelo y preparar los datasets con distintas etiquetas. Respecto al modelo español también se puede hacer un esfuerzo porque estos sean mejores y además se podría detectar el ciberbullying en internet cómo también el grado de agresividad y sentimiento de los mensajes teniendo distintas etiquetas para ello. Por último en este aspecto, se podrían construir modelos basados en otras redes sociales, por ejemplo, Facebook e Instagram.

La aplicación es mejorable respecto a su interfaz y lenguaje, utilizando tecnologías más modernas para ello. Respecto a su funcionamiento se podría implementar tomar mensajes de una fecha dada e incluso seleccionar más países para hacer un filtrado.

Respecto a los análisis se podrían tomar muchos más mensajes para ello y en el caso de la guerra de ucrania tomar muchos más tweets por cada fecha para tener un análisis completo. Por último, se podrían comparar estos con otra red social y ver otros aspectos que no hayan sido analizados.

ANEXO I – Lista de Interesados

Código	Nombre	Rol	Tipo	Comunicación	Correo corporativo	Expectativas sobre el interesado	Interés	Estrategia de gestión
IN-01	Francisco Javier Ortega Rodríguez	Product Owner, Mentor	Interno	Correo y reuniones (Tutorías)	javierortega@us.es	Recibir conocimientos y feedback para el desarrollo del proyecto	Cumplimiento de los objetivos del proyecto y cumplimiento de los hitos marcados	Reuniones periódicas en las diferentes fases del desarrollo para resolver dudas y ofrecer mejoras durante el ciclo de vida del proyecto
IN-02	Ezequiel Pérez Sosa	Participante del equipo de desarrollo	Interno	Correo y reuniones	ezepersos@alum.us.es	Cumplimiento de las responsabilidades acordadas en el acta de constitución y consecución de los objetivos marcados en la planificación del proyecto	Adquirir y poner en práctica conocimientos durante el proceso de desarrollo del proyecto, cumplimentando y consiguiendo los hitos	Monitorizar la comunicación y colaboración de las revisiones del proyecto, así como favorecer los horarios para el retorno de feedback durante el desarrollo

Código	Nombre	Rol	Tipo	Comunicación	Correo corporativo	Expectativas sobre el interesado	Interés	Estrategia de gestión
							propuestos en el alcance de este	
IN-03	Mercado de productos de consumo y servicios	Clientes	Externo	Correo	-	Conocer cómo se recibe un producto o servicio por el público monitoreando las redes sociales	Obtener servicio del proyecto cuando esté finalizado	Analizar el mercado para recibir feedback sobre sus productos o servicios
IN-04	Investigadores sociales	Clientes	Externo	Correo	-	Analizar y recopilar información haciendo uso del servicio que ofrece el proyecto	Obtener servicio del proyecto cuando esté finalizado	Analizar las redes sociales y el comportamiento humano en estas

Tabla 28 - Lista de interesado

ANEXO II – Glosario de Términos

Código	Término	Definición
T01	4Chan	4Chan es un foro lanzado originalmente el 1 de octubre de 2003. Generalmente se publica contenido de forma anónima y está ligado al activismo en internet. (4Chan - Wikipedia, s.f.)
T02	ADAM	Adam es un algoritmo de optimización de redes neuronales ampliamente utilizado.
T03	Análisis de datos	El análisis de datos es un proceso que consiste en inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil, para sugerir conclusiones y apoyo en la toma de decisiones. (Análisis de datos - Wikipedia, s.f.)
T04	Analista	El analista de negocio es la persona que posee conocimientos técnicos sobre la construcción de sistemas informáticos y al mismo tiempo comprende y está al corriente de las necesidades del usuario que requiere de esos sistemas para realizar su trabajo. (https://es.wikipedia.org/wiki/Analista_de_negocio , s.f.)
T05	API	La interfaz de programación de aplicaciones, conocida también por la sigla API, en inglés, application programming interface, es un conjunto de subrutinas, funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizada por otro software como una capa de abstracción. (API - Wikipedia, s.f.)
T06	Asalto al capitolio	El asalto al Capitolio de los Estados Unidos fue un acontecimiento que se produjo el 6 de enero de 2021 cuando partidarios del entonces presidente saliente de los Estados Unidos, Donald Trump, irrumpieron en la sede del Congreso violando la seguridad y ocupando partes del edificio durante varias horas. (Asalto al capitolio de los EE.UU. - Wikipedia, s.f.)
T07	Backend	Backend hace referencia a la parte lógica de un sistema informático.
T08	Backpropagation	Backpropagation es un método para el cálculo de gradiente para el entrenamiento de redes neuronales. (Wikipedia, Propagación hacia atrás, s.f.)
T09	Bag of words	Bag of Words o Bolsa de palabras es una forma de representación usada en PLN
T10	Batch (Entrenamiento redes)	El batch hace referencia al número de muestras que se usan en cada interacción del aprendizaje.
T11	BERT	BERT (Bidirectional Encoder Representations from Transformers) o Representación de Codificador Bidireccional de Transformadores es una técnica basada en redes neuronales para el pre-entrenamiento del procesamiento del lenguaje natural (PLN) desarrollada por Google. (BERT - Wikipedia, s.f.)

Código	Término	Definición
T12	Case/Uncased (para modelos)	Case o Uncased se utiliza para indicar si un modelo es sensible a puntuaciones y mayúsculas y minúsculas.
T13	Ciclo de vida del proyecto	El ciclo de vida del proyecto comprende la fase de desarrollo desde que se especifica el concepto de este hasta su finalización y pase a producción.
T14	Conjunto de entrenamiento (Redes neuronales)	Se trata del conjunto de muestras que se utilizan en la fase de entrenamiento de la red en la primera fase. Este conjunto es el mayor en número de elementos.
T15	Conjunto de validación (Redes neuronales)	Se trata del conjunto de muestras que se utilizan en la fase de entrenamiento de la red después del conjunto de entrenamiento. Este es el conjunto utilizado para comprobar el error.
T16	CPU	Es el componente encargado de interpretar instrucciones. Sus siglas vienen de “unidad central de procesamiento”.
T17	Crisis covid-19	Crisis covid-19 hace referencia a la crisis provocada por la pandemia de SARS-CoV-2 que afectó a todos los países a nivel global y puso en jaque a las instituciones sanitarias de todo el mundo.
T18	CSV	Se trata de un tipo de formato sencillo para presentar datos en tablas con un separador, normalmente por comas o puntos.
T19	Curva ROC	Se trata de una señal sensible a la especificidad de un sistema binario de clasificación. (Wikipedia, s.f.)
T20	Cyberbullying	Se trata del acoso que se produce hacia una persona en el ámbito de las tecnologías informáticas o redes sociales.
T21	Dataset	Un conjunto de datos (conocido también por el anglicismo dataset, comúnmente utilizado en algunos países hispanohablantes) es una colección de datos habitualmente tabulada. En el caso de datos tabulados, un conjunto de datos contiene los valores para cada una de las variables organizadas como columnas (Wikipedia, Wikipedia - Conjunto de datos, s.f.)
T22	Deep Learning	Se trata de los algoritmos que trabajan de forma automatizada entrenándose para realizar una tarea. Forma parte del campo del machine learning.
T23	Desarrollador	Un desarrollador es un programador o una compañía comercial que se dedica a uno o más aspectos del proceso de desarrollo de software. (Desarrollador Software - Wikipedia, s.f.)
T24	Discursos de odio	El discurso de odio (en inglés: hate speech) es la acción comunicativa que tiene como objetivo promover y alimentar un dogma, cargado de connotaciones discriminatorias, que atenta contra la dignidad de un grupo de individuos. (Discursos de odio - Wikipedia, s.f.)
T25	Diseñador gráfico	El diseñador gráfico es aquella persona que transmite mensajes e ideas a través de la creatividad y el pensamiento lateral. En resumidas cuentas, se encarga de transmitir ideas a través de la comunicación gráfica.

Código	Término	Definición
T26	Early stopping	Se trata de una llamada en el entrenamiento de modelos para que este pare de entrenar si la métrica a seguir está empeorando desde hace ciertas etapas para así ahorrar recursos y ajustar el modelo según los resultados obtenidos.
T27	Elecciones generales	Se trata de un proceso institucional que, a través del voto, los ciudadanos se encargan de decidir quienes ostentarán los cargos políticos dentro del marco de una democracia representativa.
T28	Feedback o retroalimentación	La realimentación ¹ —también referida de forma común como retroalimentación— es un mecanismo por el cual una cierta proporción de la salida de un sistema se redirige a la entrada, con señales de controlar su comportamiento. (Retroalimentación - Wikipedia, s.f.)
T29	GitHub	Se trata de un servicio de control de versiones. Es un servicio apoyado en git.
T30	GPU	Se trata del componente encargado del apartado gráfico. Sus siglas provienen de “unidad de procesamiento gráfico”.
T31	Hiper parámetros	Se trata de variables que hacen referencia a la configuración de entrenamiento de los modelos.
T32	HuggingFace	Se trata de una empresa emergente sobre PLN centrada en las bibliotecas de Transformers.
T33	Investigadores sociales	Se trata de personas que recopilan información de sociedad con el fin de ofrecer un servicio o producto.
T34	Jefe de proyecto	Un gestor de proyecto, también conocido con el término gerente de proyecto, director de proyecto, líder de proyecto o encargado de proyecto, es la persona que tiene la responsabilidad total del planeamiento y la ejecución acertada de cualquier proyecto. Este título se utiliza en la industria de la construcción, la arquitectura, el desarrollo de software y en diversas ocupaciones que se basan en la generación o manutención de un producto. (Gestor de proyecto - Wikipedia, s.f.)
T35	Marco temporal	Llamamos marco temporal como el periodo de tiempo desde inicio a fin en el que sucede un hecho.
T36	Kaggle	Se trata de un sitio web donde se alojan modelos, datasets, códigos, etc. Del campo de la inteligencia artificial.
T37	Librerías (Informática)	Se trata de un conjunto de implementaciones destinadas a una funcionalidad. No son utilizadas por si solas sino que se utilizan cómo apoyo para una funcionalidad concreta.
T38	Machine Learning	Se trata de la rama de la inteligencia artificial que cubre los algoritmos de aprendizaje de forma automatizada.
T39	máscara de atención (Deep learning)	Se trata de una matriz que indica al algoritmo cuáles son los elementos que no son usados como relleno.
T40	Material Fungible	Se considera material fungible como aquellos productos que se consumen cuando son utilizados sin poder ser reutilizados.
T41	Material No Fungible	Se considera material no fungible como aquellos productos que no se consumen cuando son utilizados y pueden ser reutilizados.

Código	Término	Definición
T42	Matriz de confusión	Se trata de una herramienta para medir el trabajo de un algoritmo enfrentando los aciertos y errores de cada una de las clasificaciones.
T43	Memoria RAM	Se trata de un componente encargado de almacenar información volátil y no permanente.
T44	Mentor	Persona con más experiencia a otra que ayuda brindando el conocimiento que posee.
T45	Mercado de productos de consumo y servicios	Representa el mercado donde se comercializan productos que están destinados a consumirse para satisfacer una necesidad.
T46	MLM (Deep learning)	Se trata de una técnica de entrenamiento con la que fue entrenado el modelo BERT. El modelo intenta predecir cierta parte de una entrada. (Sentence transformers, s.f.)
T47	Multihilos	Se trata de una programación basada en hilos, esto permite que cada uno de ellos utilice la CPU realizando distintas operaciones.
T48	NSP (Deep learning)	Se trata de una técnica de entrenamiento con la que fue entrenado el modelo BERT. EL modelo debe decidir que frase sigue a una dada, dándole para ello dos opciones.
T49	One-hot encoding	Se crea para procesar etiquetas múltiples. Se crea una columna por cada valor disponible y esta será 1 cuando el valor corresponda con esta.
T50	ONU	La Organización de las Naciones Unidas (ONU), también conocida simplemente como Naciones Unidas (NN. UU.), es la mayor organización internacional existente. Se creó para mantener la paz y seguridad internacionales, fomentar relaciones de amistad entre las naciones, lograr la cooperación internacional para solucionar problemas globales y servir de centro que armonice las acciones de las naciones. (Organización de las Naciones Unidas - Wikipedia, s.f.)
T51	Optimización	La optimización de software es el proceso de modificación de un software para hacer que algún aspecto de este funcione de manera más eficiente y/o utilizar menos recursos (mayor rendimiento). En general, un programa puede ser optimizado para que se ejecute más rápidamente, o sea capaz de operar con menos memoria u otros recursos, o consuman menos energía. (Optimización de software - Wikipedia, s.f.)
T52	Padding (Informática)	Se trata de una técnica de relleno. Consiste en rellenar un espacio vacío con números (normalmente todos 0 o 1).
T53	Procesado de datos	El procesamiento de datos es, en general, "la acumulación y manipulación de elementos de datos para producir información significativa." (Procesamiento de datos - Wikipedia, s.f.)
T54	Procesamiento del Lenguaje Natural (PLN)	El procesamiento de lenguaje natural, ¹ abreviado PLN —en inglés, natural language processing, NLP— es un campo de las ciencias de la computación, de la inteligencia artificial y de la lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. Se ocupa de la formulación e investigación de mecanismos eficaces

Código	Término	Definición
		computacionalmente para la comunicación entre personas y máquinas por medio del lenguaje natural, es decir, de las lenguas del mundo. (Procesamiento de lenguajes naturales - Wikipedia, s.f.)
T55	Reddit	Reddit (estilizado en minúscula como reddit) es un sitio web de marcadores sociales y agregador de noticias donde los usuarios pueden añadir textos, imágenes, videos o enlaces. (Reddit - Wikipedia, s.f.)
T56	Redes neuronales recurrentes (RNN)	Se trata de un tipo de red neuronal con pesos y una función de activación. Estas redes tienen en cuenta la salida anterior para la entrada que se está procesando. (Wikipedia, s.f.)
T57	Redes sociales	Se trata de una estructura formada en internet en las cuales los individuos se relacionan entre sí a partir de intereses comunes. Permiten el contacto entre las personas usándose como un medio de intercambio de información.
T58	Salario bruto	Se trata del salario el cual aún no ha sufrido las retenciones y cotizaciones de la nómina.
T59	SGD	SGD es un tipo de algoritmo de optimización de redes neuronales.
T60	Sigmoid	Es un tipo de función matemática usada para describir la evolución de una curva de aprendizaje. (Wikipedia, s.f.)
T61	softmax	Es un tipo de función matemática basada de cálculo de probabilidades.
T62	Sprint	Se trata de un ciclo o una iteración dentro del ciclo de vida de un proyecto bajo la metodología Scrum.
T63	Streaming	Hace referencia a la distribución de datos a través de la red utilizado a la vez que se obtiene.
T64	SWIFT	(Society for Worldwide Interbank Financial Telecommunication) Se trata de la red de comunicación financiera de bancos.
T65	TensorFlow	TensorFlow es una biblioteca de código abierto para aprendizaje automático a través de un rango de tareas, y desarrollado por Google para satisfacer sus necesidades de sistemas capaces de construir y entrenar redes neuronales para detectar y descifrar patrones y correlaciones, análogos al aprendizaje y razonamiento usados por los humanos. (TensorFlow - Wikipedia, s.f.)
T66	Testeo	Las pruebas de software (en inglés software testing) son las investigaciones empíricas y técnicas cuyo objetivo es proporcionar información objetiva e independiente sobre la calidad del producto a la parte interesada o stakeholder. Es una actividad más en el proceso de control de calidad. (Pruebas de software - Wikipedia, s.f.)
T67	Twitter	La red permite enviar mensajes de texto plano de corta longitud, con un máximo de 280 caracteres (originalmente 140), llamados tuits o tweets (aunque esta última acepción no está recogida en la RAE), que se muestran en la página principal del usuario. Los usuarios pueden suscribirse a los tweets de otros usuarios —a esto se le llama seguir y a los

Código	Término	Definición
		usuarios abonados se les llama seguidores. (Twitter - Wikipedia, s.f.)
T68	Tkinter	Se trata de una librería de Python para crear interfaces.
T69	tokenización	Se trata de la transformación de datos en bruto por símbolos o números para que sean entendidos por un sistema informático.
T70	Transformers (Machine learning)	Se trata de un modelo de Deep learning basado en los modelos de atención.
T71	Twitter	Se trata de una red social donde los usuarios publican mensajes de no más de 256 caracteres.
T72	Word embedding	Se trata de un modelo de representación de palabras basado en vectores.
T73	Word2Vec	Se trata de una técnica para procesar el lenguaje preparado para aprender asociaciones de palabras. (Wikipedia, s.f.)

Tabla 29 - Glosario de términos

ANEXO III – Bibliografía y Referencias

- BERT- Wikipedia*. (s.f.). Obtenido de [https://es.wikipedia.org/wiki/BERT_\(modelo_de_lenguaje\)](https://es.wikipedia.org/wiki/BERT_(modelo_de_lenguaje))
- 4Chan - Wikipedia*. (s.f.). Obtenido de https://es.wikipedia.org/wiki/4chan#cite_note-TakingRick-4
- Análisis de datos - Wikipedia*. (s.f.). Obtenido de https://es.wikipedia.org/wiki/An%C3%A1lisis_de_datos
- API - Wikipedia*. (s.f.). Obtenido de https://es.wikipedia.org/wiki/Interfaz_de_programaci%C3%B3n_de_aplicaciones
- Arciniegas, Y. (19 de Diciembre de 2020). *Trump acusa a China del ciberataque contra su Gobierno y lo vincula a supuesto fraude electoral*. Obtenido de <https://www.france24.com/es/ee-uu-y-canad%C3%A1/20201219-pompeo-silencio-administracion-trump-acusacion-rusia-ciberataque>
- Asalto al capitolio de los EE.UU. - Wikipedia*. (s.f.). Obtenido de https://es.wikipedia.org/wiki/Asalto_al_Capitolio_de_los_Estados_Unidos_de_2021
- Barbieri, F. a.-C.-A. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv*, 7.
- Bengio, D. B. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv*.
- Canuma, P. (s.f.). <https://medium.datadriveninvestor.com/the-brief-history-of-nlp-c90f331b6ad7>. Obtenido de Medium - The brief history of NLP.
- CNN. (2020). *PRESIDENTIAL RESULTS*. Obtenido de <https://edition.cnn.com/election/2020/results/president#mapmode=lead>
- CNN. (Mayo de 2022). *Así ha sido, día a día, la guerra en Ucrania: datos y cronología sobre la invasión rusa*. Obtenido de <https://cnnespanol.cnn.com/2022/06/04/guerra-ucrania-cronologia-orix/>
- Deng, K. D. (2015). Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 17-27.
- Desarrollador Software - Wikipedia*. (s.f.). Obtenido de Un desarrollador es un programador o una compañía comercial que se dedica a uno o más aspectos del proceso de desarrollo de software.
- Devlin, J. (2018). *Google*. Obtenido de Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing: <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>

digital, M. M.-P. (Abril de 2021). *Youtube*. Obtenido de Procesamiento del Lenguaje Natural (PLN): De las redes shallow a los transformers.:
<https://www.youtube.com/watch?v=glhRZlieZ3g>

Discursos de odio - Wikipedia. (s.f.). Obtenido de
https://es.wikipedia.org/wiki/Discurso_de_odio

EuropaPress. (29 de Julio de 2021). *Rusia rechaza las acusaciones de Biden sobre un supuesto plan para interferir en las elecciones de 2022*. Obtenido de
<https://www.europapress.es/internacional/noticia-rusia-rechaza-acusaciones-biden-supuesto-plan-interferir-elecciones-2022-20210728154436.html>

Face, H. (s.f.). *Hugging Face - bert-base-uncased*. Obtenido de Hugging Face :
<https://huggingface.co/bert-base-uncased?text=Paris+is+the+%5BMASK%5D+of+France>.

Gestor de proyecto - Wikipedia. (s.f.). Obtenido de
https://es.wikipedia.org/wiki/Gestor_de_proyecto#:~:text=Un%20gestor%20de%20proyecto%2C%20tambi%C3%A9n,ejecuci%C3%B3n%20acertada%20de%20cualquier%20proyecto.

Gil, T. (1 de Febrero de 2021). *Estados Unidos vs China: ¿puede la relación entre Pekín y Washington recuperarse tras cuatro años de Donald Trump?* Obtenido de
<https://www.bbc.com/mundo/noticias-internacional-55805132>

Gobierno EE. UU. (s.f.). *Censo poblacional*. Obtenido de <https://www.census.gov/popclock/>

Google. (s.f.). *Google Trends*. Obtenido de <https://trends.google.es/trends/>

Horev, R. (10 de 11 de 2018). *BERT Explained: State of the art language model for NLP*. Obtenido de Towards Data Science: <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>

<https://es.wikipedia.org/wiki/4chan>. (s.f.).

https://es.wikipedia.org/wiki/Analista_de_negocio. (s.f.). Obtenido de Analista - Wikipedia

IBM, G. U. (1954). Georgetown-IBM experiment.

INE. (2021). *Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares*. Obtenido de https://www.ine.es/prensa/tich_2021.pdf

InteractiveChaos. (s.f.). *Adam - InteractiveChaos*. Obtenido de
<https://interactivechaos.com/es/manual/tutorial-de-machine-learning/adam#:~:text=Adam%20o%20Adaptative%20Moment%20Optimization,tasas%20de%20aprendizaje%20por%20variable>.

InteractiveChaos. (s.f.). *InteractiveChaos - Stochastic Gradient Descent*. Obtenido de
<https://interactivechaos.com/es/manual/tutorial-de-machine-learning/stochastic-gradient-descent#:~:text=El%20optimizador%20Stochastic%20Gradient%20Descent,de%20esta%2C%20el%20gradiente%20y>

Jeremy Diamond, K. L. (21 de Mayo de 2022). *La estrategia de Biden con Corea del Norte está muy lejos de la mediática diplomacia de Trump*. Obtenido de

- <https://cnnespanol.cnn.com/2022/05/21/la-estrategia-de-biden-con-corea-del-norte-esta-muy-lejos-de-trump-trax/>
- Kaggle. (s.f.). *Kaggle - Sentiment140*. Obtenido de Kaggle:
<https://www.kaggle.com/datasets/kazanova/sentiment140>
- Keras. (s.f.). *Keras - Dropout layer*. Obtenido de
https://keras.io/api/layers/regularization_layers/dropout/
- Keras. (s.f.). *keras API - Dense layer*. Obtenido de
https://keras.io/api/layers/core_layers/dense/
- Kuutila, M. V. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, 16-32.
- Los angeles times. (4 de Marzo de 2022). *Rusia bloquea el acceso a Twitter y Facebook*. Obtenido de <https://www.latimes.com/espanol/internacional/articulo/2022-03-04/rusia-bloquea-el-acceso-a-twitter-y-facebook>
- Magdy, I. (. (2021). A comparative study of effective approaches for Arabic sentiment analysis. *Information Processing & Management*, 102438.
- Marr, B. (21 de 05 de 2018). *Forbes*. Obtenido de How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read:
<https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6110bb8360ba>
- Mishev, K. a. (2020). Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access*, 131662-131682.
- Mouthami, K. a. (2013). Sentiment analysis and classification based on textual reviews. En *2013 International Conference on Information Communication and Embedded Systems (ICICES)* (págs. 271-276).
- Omer Levy, Y. G. (2015). Improving Distributional Similarity. *MIT Press, Transactions of the Association for Computational Linguistics, Volume 3*, 226.
- Optimización de software - Wikipedia*. (s.f.). Obtenido de
https://es.wikipedia.org/wiki/Optimizaci%C3%B3n_de_software
- Organización de las Naciones Unidas - Wikipedia*. (s.f.). Obtenido de
https://es.wikipedia.org/wiki/Organizaci%C3%B3n_de_las_Naciones_Unidas
- Polosukhin, A. V. (2017). Attention Is All You Need. *arXiv*.
- Procesamiento de datos - Wikipedia*. (s.f.). Obtenido de
https://es.wikipedia.org/wiki/Procesamiento_de_datos
- Procesamiento de lenguajes naturales - Wikipedia*. (s.f.). Obtenido de
https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales
- Pruebas de software - Wikipedia*. (s.f.). Obtenido de
https://es.wikipedia.org/wiki/Pruebas_de_software
- Reddit - Wikipedia*. (s.f.). Obtenido de <https://es.wikipedia.org/wiki/Reddit>

Retroalimentación - Wikipedia. (s.f.). Obtenido de <https://es.wikipedia.org/wiki/Realimentaci%C3%B3n>

Roa, M. M. (7 de Marzo de 2022). *¿Cuáles son las redes sociales más usadas en Rusia?* Obtenido de <https://es.statista.com/grafico/26998/redes-sociales-con-el-mayor-porcentaje-de-usuarios-en-rusia/>

Sentence transformers. (s.f.). *SBERT - MLM*. Obtenido de https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

Tan, T. (s.f.). *Towards Data Science*. Obtenido de Evolution of Language Models: N-Grams, Word Embeddings, Attention & Transformers: <https://towardsdatascience.com/evolution-of-language-models-n-grams-word-embeddings-attention-transformers-a688151825d2>

TensorFlow - Wikipedia. (s.f.). Obtenido de <https://es.wikipedia.org/wiki/TensorFlow>

Toutanova, J. D.-}. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language. *CoRR*.

Turing, A. (1950). Computing machinery and intelligence. *Mind*.

Twitter - Wikipedia. (s.f.). Obtenido de <https://es.wikipedia.org/wiki/Twitter>

Twitter. (2016). *Las elecciones EE. UU. de 2020 y Twitter*. Obtenido de <https://help.twitter.com/es/using-twitter/us-elections>

Velasco, L. (26 de Abril de 2020). *Optimizadores en redes neuronales profundas: un enfoque práctico*. Obtenido de <https://velascoluis.medium.com/optimizadores-en-redes-neuronales-profundas-un-enfoque-pr%C3%A1ctico-819b39a3eb5>

veritessa. (s.f.). *StackExchange - Softmax vs Sigmoid function*. Obtenido de <https://stats.stackexchange.com/questions/233658/softmax-vs-sigmoid-function-in-logistic-classifier>

Weyrich, B. L. (2021). A survey on long short-term memory networks for time series prediction. *Procedia CIRP*, 650-655.

Wikipedia. (s.f.). Obtenido de <https://es.wikipedia.org/wiki/4chan>

Wikipedia. (s.f.). Obtenido de https://es.wikipedia.org/wiki/An%C3%A1lisis_de_datos

Wikipedia. (2022). *Anexo:Cronología de la invasión rusa de Ucrania de 2022*. Obtenido de https://es.wikipedia.org/wiki/Anexo:Cronolog%C3%ADa_de_la_invasi%C3%B3n_rusa_de_Ucrania_de_2022

Wikipedia. (s.f.). *Curva ROC*. Obtenido de https://es.wikipedia.org/wiki/Curva_ROC

Wikipedia. (s.f.). *Escándalo Facebook-Cambridge Analytica*. Obtenido de https://es.wikipedia.org/wiki/Esc%C3%A1ndalo_Facebook-Cambridge_Analytica

Wikipedia. (s.f.). *Función sigmoide*. Obtenido de https://es.wikipedia.org/wiki/Funci%C3%B3n_sigmoide

Wikipedia. (s.f.). *Propagación hacia atrás*. Obtenido de https://es.wikipedia.org/wiki/Propagaci%C3%B3n_hacia_atr%C3%A1s#:~:text=La%20p

propagaci%C3%B3n%20hacia%20atr%C3%A1s%20de,propagaci%C3%B3n%20%E2%80%93%20adaptaci%C3%B3n%20de%20dos%20fases.

Wikipedia. (s.f.). *Red neuronal recurrente*. Obtenido de https://es.wikipedia.org/wiki/Red_neuronal_recurrente

Wikipedia. (s.f.). *Wikipedia - Conjunto de datos*. Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Conjunto_de_datos

Wikipedia. (s.f.). *Word2vec*. Obtenido de <https://es.wikipedia.org/wiki/Word2vec>

Winograd, T. (1970). SHRDLU.

Wolf, V. S. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*.

Yang, K. (2021). Transformer-based Korean Pretrained Language Models: A Survey on Three Years of Progress. *arXiv*.