

Perfilado automático de usuarios en corpus sociales sobre el movimiento *Black Lives Matter*

Nicolás Míguez García

Grado en Ingeniería Informática (Mención en Computación)

Patricia Martín Rodilla
David Otero Freijeiro



facultade de
informática
da coruña



UNIVERSIDADE DA CORUÑA

Índice general

- 1 Introducción
- 2 Fundamentos
 - Estado del arte
 - Algoritmos de perfilado
- 3 Fenómeno #BLM
 - Contexto movimiento
 - Demo
 - Análisis resultados
- 4 Metodología y gestión del proyecto
- 5 Diseño
- 6 Conclusiones

Introducción

Author profiling

El **author profiling** consiste en el uso del NLP y aprendizaje automático para la extracción automática de características de autores de textos.

Introducción

Author profiling

El **author profiling** consiste en el uso del NLP y aprendizaje automático para la extracción automática de características de autores de textos. Permite el análisis masivo de datos de redes sociales para:

El **author profiling** consiste en el uso del NLP y aprendizaje automático para la extracción automática de características de autores de textos.

Permite el análisis masivo de datos de redes sociales para:

- Investigación sociológica.
- Política.
- Marketing.

- Estudio del **estado del arte** del *author profiling* en materia de género y edad sobre textos en español.
- Reproducción de los **algoritmos** con mejor rendimiento.
- Construcción de una **herramienta** para el análisis de grandes conjuntos de usuarios.
- Estudio de **resultados** de perfilado sobre colección...

Índice general

- 1 Introducción
- 2 Fundamentos
 - Estado del arte
 - Algoritmos de perfilado
- 3 Fenómeno #BLM
 - Contexto movimiento
 - Demo
 - Análisis resultados
- 4 Metodología y gestión del proyecto
- 5 Diseño
- 6 Conclusiones

- Trabajos tempranos sobre blogs (2006 y 2007).

- Trabajos tempranos sobre blogs (2006 y 2007).
- Competiciones PAN a partir de 2013.

- Trabajos tempranos sobre blogs (2006 y 2007).
- Competiciones PAN a partir de 2013.
- IberLEF 2022.

- Trabajos tempranos sobre blogs (2006 y 2007).
- Competiciones PAN a partir de 2013.
- IberLEF 2022.
- Recursos limitados en idioma español.

- Trabajos tempranos sobre blogs (2006 y 2007).
- Competiciones PAN a partir de 2013.
- IberLEF 2022.
- Recursos limitados en idioma español.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

2º Grivas:

- 3º en AP-PAN 2015.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

2º Grivas:

- 3º en AP-PAN 2015.
- El mejor en edad.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

2º Grivas:

- 3º en AP-PAN 2015.
- El mejor en edad.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

3ª Modaresi:

- 2º en PAN 2016.

2º Grivas:

- 3º en AP-PAN 2015.
- El mejor en edad.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

2º Grivas:

- 3º en AP-PAN 2015.
- El mejor en edad.

3ª Modaresi:

- 2º en PAN 2016.
- Independencia del género escritura.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

2º Grivas:

- 3º en AP-PAN 2015.
- El mejor en edad.

3ª Modaresi:

- 2º en PAN 2016.
- Independencia del género escritura.
- Mejor en clasificación de género usuarios.

Se seleccionaron 3 algoritmos:

1º Carrasco y Rosillo

- 1º IberLEF 2022.
- Más prometedor.
- Peor rendimiento de los tres.
- No se ha terminado usando.

2º Grivas:

- 3º en AP-PAN 2015.
- El mejor en edad.

3ª Modaresi:

- 2º en PAN 2016.
- Independencia del género escritura.
- Mejor en clasificación de género usuarios.

Índice general

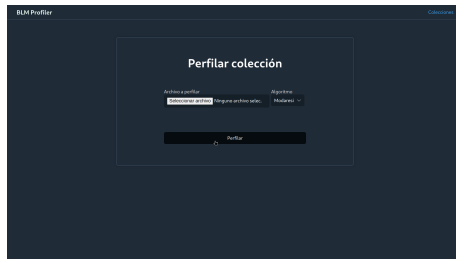
- 1 Introducción
- 2 Fundamentos
 - Estado del arte
 - Algoritmos de perfilado
- 3 Fenómeno #BLM
 - Contexto movimiento
 - Demo
 - Análisis resultados
- 4 Metodología y gestión del proyecto
- 5 Diseño
- 6 Conclusiones

The image shows the Black Lives Matter logo, which consists of the words "BLACK", "LIVES", and "MATTER" stacked vertically in a bold, black, sans-serif font. The word "LIVES" is enclosed within a black rectangular box, while "BLACK" and "MATTER" are placed on a solid yellow background.

The logo for Black Lives Matter is displayed on a bright yellow rectangular background. The text is in a bold, black, sans-serif font. The word "BLACK" is at the top, "LIVES" is in the middle and enclosed within a black rectangular box, and "MATTER" is at the bottom.

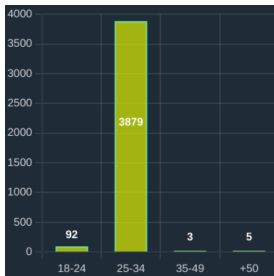
Varios investigadores construyeron una colección para el estudio de este movimiento:

- Un año de actividad.
- +260.000 publicaciones de +90.000 usuarios.
- Inglés y español.

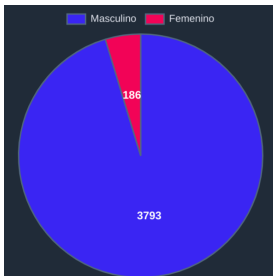


Resultados perfilado

Algoritmo de Grivas



(a) Edad

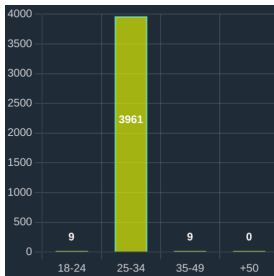


(b) Género

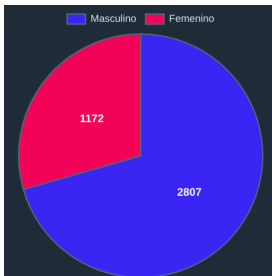
Género \ Edad	Edad				Total
	18-24	25-34	35-49	+50	
Femenino	31	154	0	1	186
Masculino	61	3725	3	4	3793
Total	92	3879	3	5	3979

Resultados perfilado

Algoritmo de Modaresi



(a) Edad



(b) Género

Género \ Edad	Edad				Total
	18-24	25-34	35-49	+50	
Femenino	5	1162	5	0	1172
Masculino	4	2799	4	0	2807
Total	9	3961	9	0	3979

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años $> 95\%$ (desequilibrio entrenamiento)

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años > 95 % (desequilibrio entrenamiento)

Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1

Fuente: Pew Research Center (estudio 2016).

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años > 95 % (desequilibrio entrenamiento)
- Usuarios masculinos > 75 %.

Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1

Fuente: Pew Research Center (estudio 2016).

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años > 95 % (desequilibrio entrenamiento)
- Usuarios masculinos > 75 %.
- Distribución Reddit \implies distribución corpus.

Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1

Fuente: Pew Research Center (estudio 2016).

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años > 95 % (desequilibrio entrenamiento)
- Usuarios masculinos > 75 %.
- Distribución Reddit \implies distribución corpus.
- 48.98 % \rightarrow 1 publicación.
80 % \rightarrow menos de 5.



Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1

Fuente: Pew Research Center (estudio 2016).

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años > 95 % (desequilibrio entrenamiento)
- Usuarios masculinos > 75 %.
- Distribución Reddit \implies distribución corpus.
- 48.98 % \rightarrow 1 publicación.
80 % \rightarrow menos de 5.
↓
- Baja fiabilidad.

Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1

Fuente: Pew Research Center (estudio 2016).

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años > 95 % (desequilibrio entrenamiento)
- Usuarios masculinos > 75 %.
- Distribución Reddit \implies distribución corpus.
- 48.98 % \rightarrow 1 publicación.
80 % \rightarrow menos de 5.



- Baja fiabilidad.
- Resultados similares inglés.

Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1

Fuente: Pew Research Center (estudio 2016).

Resultados perfilado

Discusión y conclusiones

- Grupo 25-34 años > 95 % (desequilibrio entrenamiento)
- Usuarios masculinos > 75 %.
- Distribución Reddit \implies distribución corpus.
- 48.98 % \rightarrow 1 publicación.
80 % \rightarrow menos de 5.
↓
- Baja fiabilidad.
- Resultados similares inglés.
- Conclusiones #BLM.

Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1

Fuente: Pew Research Center (estudio 2016).

Índice general

- 1 Introducción
- 2 Fundamentos
 - Estado del arte
 - Algoritmos de perfilado
- 3 Fenómeno #BLM
 - Contexto movimiento
 - Demo
 - Análisis resultados
- 4 Metodología y gestión del proyecto
- 5 Diseño
- 6 Conclusiones

Factores considerados:

- Carácter innovador.
- Escaso conocimiento.

Factores considerados:

- Carácter innovador.
- Escaso conocimiento.



Scrum

- Metodología ágil.
- Transparencia, inspección y adaptación.

Factores considerados:

- Carácter innovador.
- Escaso conocimiento.



Scrum

- Metodología ágil.
- Transparencia, inspección y adaptación.

Adaptaciones Scrum

Factores considerados:

- Carácter innovador.
- Escaso conocimiento.



Scrum

- Metodología ágil.
- Transparencia, inspección y adaptación.

Adaptaciones Scrum

- Roles
- Eventos
- Artefactos

- 9 sprints en total.

Gestión del proyecto

Estimación y costes

- 9 sprints en total.
- Cada uno \approx 45 horas de trabajo (3 h/día).

Gestión del proyecto

Estimación y costes

- 9 sprints en total.
- Cada uno \approx 45 horas de trabajo (3 h/día).
- Recursos humanos: alumno (18€/h) y directores (31€/h).

- 9 sprints en total.
- Cada uno ≈ 45 horas de trabajo (3 h/día).
- Recursos humanos: alumno (18€/h) y directores (31€/h).
- Ordenador portátil ≈ 428 €.

Gestión del proyecto

Estimación y costes

- 9 sprints en total.
- Cada uno ≈ 45 horas de trabajo (3 h/día).
- Recursos humanos: alumno (18€/h) y directores (31€/h).
- Ordenador portátil ≈ 428 €.

Cálculo final:

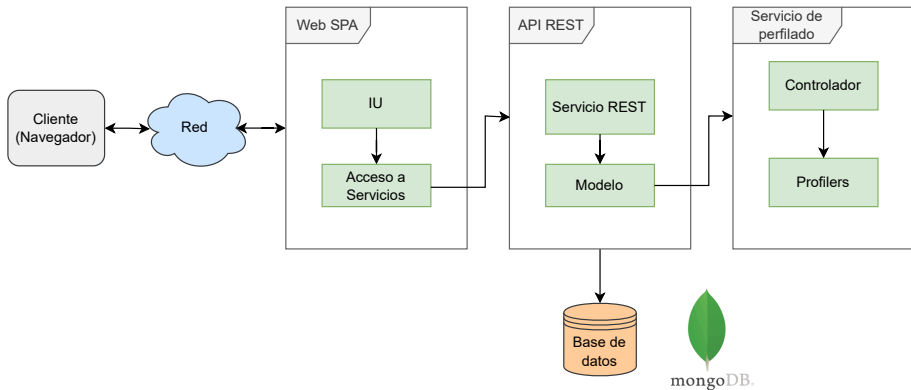
Rol	Coste/hora	Tiempo de trabajo	Total
Equipo	18€	45h x 9 <i>sprints</i>	7.290 €
<i>Project Managers</i>	31€	2 x 1.5h x 9 <i>sprints</i>	837 €
Material	—	—	428 €
Total	—	—	8.555 €

Índice general

- 1 Introducción
- 2 Fundamentos
 - Estado del arte
 - Algoritmos de perfilado
- 3 Fenómeno #BLM
 - Contexto movimiento
 - Demo
 - Análisis resultados
- 4 Metodología y gestión del proyecto
- 5 Diseño**
- 6 Conclusiones

Se ha optado por una arquitectura cliente-servidor distribuida en 3 capas.

Se ha optado por una arquitectura cliente-servidor distribuida en 3 capas.



Índice general

- 1 Introducción
- 2 Fundamentos
 - Estado del arte
 - Algoritmos de perfilado
- 3 Fenómeno #BLM
 - Contexto movimiento
 - Demo
 - Análisis resultados
- 4 Metodología y gestión del proyecto
- 5 Diseño
- 6 Conclusiones

Conclusiones

Evaluación estado actual y lecciones aprendidas

Resultados:

- Varios algoritmos de perfilado.

Conclusiones

Evaluación estado actual y lecciones aprendidas

Resultados:

- Varios algoritmos de perfilado.
- Herramienta para el análisis de grandes colecciones.

Conclusiones

Evaluación estado actual y lecciones aprendidas

Resultados:

- Varios algoritmos de perfilado.
- Herramienta para el análisis de grandes colecciones.
- Análisis movimiento #BLM.

Conclusiones

Evaluación estado actual y lecciones aprendidas

Resultados:

- Varios algoritmos de perfilado.
- Herramienta para el análisis de grandes colecciones.
- Análisis movimiento #BLM.

Lecciones aprendidas:

Conclusiones

Evaluación estado actual y lecciones aprendidas

Resultados:

- Varios algoritmos de perfilado.
- Herramienta para el análisis de grandes colecciones.
- Análisis movimiento #BLM.

Lecciones aprendidas:

- Importancia organización metodológica.

Conclusiones

Evaluación estado actual y lecciones aprendidas

Resultados:

- Varios algoritmos de perfilado.
- Herramienta para el análisis de grandes colecciones.
- Análisis movimiento #BLM.

Lecciones aprendidas:

- Importancia organización metodológica.
- Relación conocimientos.

Conclusiones

Evaluación estado actual y lecciones aprendidas

Resultados:

- Varios algoritmos de perfilado.
- Herramienta para el análisis de grandes colecciones.
- Análisis movimiento #BLM.

Lecciones aprendidas:

- Importancia organización metodológica.
- Relación conocimientos.
- Crecimiento personal.

Algoritmos perfilado:

- Mejora corpus entrenamiento.

Algoritmos perfilado:

- Mejora corpus entrenamiento.
- Nuevos modelos (LLM).

Algoritmos perfilado:

- Mejora corpus entrenamiento.
- Nuevos modelos (LLM).
- Aumento de atributos.

Funcionalidad aplicación:

- Perfilado asíncrono.

Algoritmos perfilado:

- Mejora corpus entrenamiento.
- Nuevos modelos (LLM).
- Aumento de atributos.

Funcionalidad aplicación:

- Perfilado asíncrono.
- Exportación de datos.

Algoritmos perfilado:

- Mejora corpus entrenamiento.
- Nuevos modelos (LLM).
- Aumento de atributos.

Funcionalidad aplicación:

- Perfilado asíncrono.
- Exportación de datos.
- Explicación de predicción.

¡Gracias por su atención!