

Análisis de Sentimientos en Tweets sobre el entrenador del equipo de fútbol y su Relación con el rendimiento deportivo: Estimación del rendimiento de los Entrenadores

Pablo Rodríguez Suárez

1. Introducción

1.1 Contexto y Motivación

En el siglo XXI las redes sociales han cambiado totalmente la forma de seguir el deporte. Hasta tal punto que hoy en día se busca la polarización. Un ejemplo de ello es el aumento del dominio del binomio Barcelona-Real Madrid en la Liga Española tras la llegada de las redes sociales(y Cristiano y Messi quizás también). Al inicio se realizaban opiniones en foros locales, pero ahora todo se unificó en la plataforma Twitter donde cada uno puede debatir con cualquier persona del mundo. Esta avalancha de sentimientos en la plataforma puede tener un gran impacto en las decisiones deportivas y administrativas del club, ya que no solo es el lugar para expresar tu opinión, sino también para leer las de los demás.

Con la gran cantidad de dinero que hay hoy en día en juego de patrocinadores, los equipos no deben centrarse solamente en los resultados deportivos para decidir la continuidad del entrenador, sino también de la opinión pública. Pese a tener un buen proyecto, una racha negativa junto a un sentimiento muy negativo en las redes sociales(falta de carisma del entrenador por ejemplo) puede ejercer presión en la directiva del club para cesar al entrenador.

Este proyecto busca analizar la reacción de los seguidores en la plataforma de Twitter respecto a los resultados deportivos y qué opiniones tienen un mayor peso sobre la sociedad.

Los clubes deportivos no solo deben considerar los resultados en el campo, sino también la opinión pública. Un sentimiento negativo persistente en las redes sociales, combinado con una serie de malos resultados, puede ejercer presión sobre la directiva de un club para que considere el despido de un entrenador. Este proyecto busca analizar y entender cómo los sentimientos expresados en Twitter después de los partidos, junto con los resultados deportivos, pueden predecir si un entrenador será despedido.

1.1 Motivación

Mi formación académica es el grado de Física y con más de dos años de experiencia de investigación en proyectos de magnetismo. Tras tener la experiencia de trabajar con los mejores investigadores en el marco europeo decido entrar en el campo de Data Science. Pero surge la siguiente pregunta: **¿Un físico puede ser Data Scientist?**

La respuesta es sí. Muchos de los modelos desarrollados para describir fenómenos físicos, son aplicados posteriormente en otros campos.

1.3 Objetivos del Proyecto

El objetivo principal de este proyecto es desarrollar un modelo que utilice el análisis de sentimientos de tweets y los resultados deportivos para predecir el despido de entrenadores en clubes de fútbol. Los objetivos específicos incluyen:

- Recolectar y analizar publicaciones de Twitter tras el resultado deportivo.
- Aplicar técnicas de análisis de sentimientos a los tweets recolectados para determinar el estado de ánimo de los aficionados.
- Correlacionar los sentimientos expresados y la actividad en la red social con los resultados de los partidos.
- Desarrollar una base de datos que permita determinar el rendimiento y la situación laboral de un entrenador.

Un objetivo mucho más ambicioso pero que se sale de las expectativas de este trabajo académico, sería realizar un estudio de las opiniones tras la jornada para poder predecir con qué tipo de publicaciones(o cantidad de ellas) puedes influir de forma más positiva o negativa en la opinión de los seguidores.

1.3 Importancia del Estudio

Este estudio es importante porque proporciona una nueva dimensión al análisis deportivo, combinando el rendimiento en el campo con la percepción pública expresada en las redes sociales. Un nuevo campo que se encuentra en auge y cuyos conocimientos se pueden aplicar a otros muchos casos donde hay mucho dinero en juego como pueden ser unas elecciones. Al predecir con mayor precisión cuándo un entrenador podría ser despedido, los clubes pueden tomar decisiones más informadas, gestionando mejor las expectativas de los aficionados y minimizando las sorpresas y las consecuencias de un cambio de entrenador. Además de entender el funcionamiento de esta plataforma, también podrían usarlo a su favor para acelerar o retrasar la decisión.

2. Recolección y Preparación de Datos

2.1 Fuente de Datos

El proyecto se basa en dos fuentes principales de datos:

1. **Tweets de Twitter:** Se obtuvieron los tweets “a mano” en los que se mencionan al entrenador del equipo lo cuál llevó una gran cantidad de tiempo. Sin embargo, en mi opinión es casi mejor para asegurarte tener una buena calidad de datos. Se decidieron estudiar los equipos de segunda para tener una mayor calidad de los datos obtenidos.
2. **Resultados Deportivos:** Los resultados de los partidos se obtuvieron de la página rdf. Esta información incluye detalles del encuentro como el equipo local, el equipo visitante, los goles anotados por cada equipo y la fecha del partido. También se obtuvo la clasificación al finalizar la jornada para obtener información sobre la situación actual del equipo.
3. **Base datos MongoDB.** Pese a que la información se encuentre en archivos formato excel. Hay un código para pasar todos estos archivos automáticamente a una base de datos de MongoDB.

2.2 Descripción de los Datos

Los datos recolectados de Twitter incluyen:

- **Contenido de los Tweets:** El texto de los tweets que mencionan al entrenador después de un partido.
- **Fecha:** El día en que se publicó cada tweet.
- **Usuario:** Se podría determinar también qué usuarios tienen un mayor peso en la opinión pública.

Los datos deportivos incluyen para los encuentros:

- **Goles Anotados:** Número de goles anotados por el equipo local y el equipo visitante.
- **Fecha del partido:** Día y hora en que se jugó cada partido.

Y para las clasificaciones:

- **Posición actual en la tabla**
- **Distancia en posiciones con el descenso.**
- **Si el equipo se encuentra en descenso**(buscando un mayor peso en la variable para decidir el despido)
- **Si el equipo se encuentra en playoff.**
- **Diferencia de posiciones respecto al año anterior**(forma indirecta de estimar si el equipo aspira a estar en la parte alta de la tabla o busca la salvación).
- **La racha en los últimos cuatro partidos**

2.3 Limpieza y Preprocesamiento

El proceso de limpieza y preprocesamiento de los datos es crucial para asegurar que el análisis posterior sea preciso y significativo. En el caso de los tweets, se realizaron los siguientes pasos:

- **Eliminación de Duplicados:** Se identificaron y eliminaron tweets duplicados para evitar sesgos en el análisis.

- **Normalización del Texto:** Se aplicaron técnicas de procesamiento de lenguaje natural (NLP) para limpiar el texto, eliminando caracteres especiales, emoticonos, menciones, respuestas, URLs, y realizando lematización para unificar palabras con la misma raíz.
- **Filtrado de Fechas:** Se filtraron los tweets que no corresponden al período de tiempo relevante, centrando el análisis en los días inmediatamente posteriores a los partidos.

Para los datos deportivos:

- **Validación de Fechas y Equipos:** Se corrigieron los nombres de equipos con tilde, ya que estaban modificados tras el proceso de extracción de datos. En esta parte no se necesitó más limpieza de datos.

3. Análisis de Sentimientos

3.1 Metodología de Análisis de Sentimientos

Para evaluar el sentimiento expresado en los tweets, se emplearon técnicas de análisis de sentimientos utilizando la herramienta **VADER** (Valence Aware Dictionary and sEntiment Reasoner). VADER es una herramienta de análisis de sentimientos útil para textos cortos como tweets. A pesar de que VADER esté optimizado al lenguaje informal que es utilizado en las redes sociales, se hizo una limpieza en los tweets de emoticonos, signos de puntuación...

El proceso de análisis de sentimientos con VADER fue el siguiente:

- **Tokenización(pendiente):** Aunque VADER trabaja directamente con frases completas, los tweets se deben descomponer en palabras y frases para un análisis detallado. Una estrategia que se realizó fue analizar solo los sentimientos que son relacionados con el entrenador(puedes estar contento con la victoria pero enfadado con el planteamiento del partido)(o contento con el equipo pero enfadado con el árbitro...) con la librería re, indicando que obtenga la frase delimitada por signos de puntuaciones y conjunciones donde se menciona al entrenador. Pero debido a la falta de tiempo se decidió realizar el estudio sin la tokenización ya que habría que hacer un control de sí en todos los tweets se resumió correctamente el tweet o no.
- **Asignación de Sentimientos:** A cada tweet se le asignó una puntuación que indican el sentimiento positivo, negativo, y neutral. Además, VADER proporciona un "compound score", que es un valor numérico normalizado entre -1 (muy negativo) y 1 (muy positivo), representando la intensidad del sentimiento general del tweet. La media de tweets no suele ser superior a 0.3.
- **Promedio de Sentimientos:** Se calculó el promedio de la puntuación "compound" para todos los tweets relacionados con un partido específico. Este promedio permitió obtener una medida general del sentimiento de los aficionados después de cada partido. Se observa que para hablar del entrenador se tiende a hacer cuando la

opinión es negativa, así que los valores suelen ser negativos o neutros salvo en casos de gran euforia.

3.2 Integración de Resultados Deportivos

Una vez calculado el sentimiento general de los tweets, se procedió a correlacionarlo con los resultados deportivos de los partidos. En este análisis, se identificaron los siguientes patrones:

- **Relación entre derrotas, número de tweets y Sentimiento Negativo:** Los resultados mostraron que el número de tweets aumentaba significativamente después de una derrota, y también el sentimiento negativo. Sin embargo, salvo para rachas muy buenas, tras una victoria el sentimiento suele ser neutro y con muy poca cantidad de tweets.
- **Variación del Sentimiento en Función de la posición de la tabla respecto al año anterior:** Si el equipo va claramente mejor que el año anterior, del entrenador apenas se habla pero como no esté en los puestos esperados el sentimiento es bastante negativo.

3.3 Ejemplos y Visualizaciones

A continuación se muestran los datos obtenidos para los casos en los que los entrenadores han sido despedidos antes de la jornada 10:

Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia Posiciones	Racha	Despedido
0 13 Aug	-0.006137	19	0	12		7	0	0		3	0
1 21 Aug	-0.148300	27	-2	19		0	1	0		-4	1
2 28 Aug	0.036686	14	0	20		-1	1	0		-5	1
3 03 Sep	-0.056287	23	0	19		0	1	0		-4	2
4 09 Sep	-0.110248	29	-2	21		-2	1	0		-6	3
5 16 Sep	-0.063339	38	0	20		-1	1	0		-5	2
6 23 Sep	-0.086365	20	2	18		1	0	0		-3	3
7 01 Oct	-0.187363	35	-1	19		0	1	0		-4	5
8 04 Oct	-0.153860	60	-1	20		-1	1	0		-5	4
9 07 Oct	-0.085738	26	-1	21		-2	1	0		-6	4

ziganda Huesca 23-24

Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia Posiciones	Racha	Despedido
0 12 Aug	-0.212188	26	0	12		7	0	0		-10	0
1 20 Aug	-0.206587	31	0	13		6	0	0		-11	1
2 27 Aug	0.037635	26	2	10		9	0	0		-8	2
3 05 Sep	-0.154751	35	0	11		8	0	0		-9	5
4 10 Sep	0.000000	1	3	5		14	0	1		-3	6
5 16 Sep	-0.198215	48	-1	10		9	0	0		-8	8
6 25 Sep	-0.257293	29	0	11		8	0	0		-9	7
7 01 Oct	-0.180543	56	-2	12		7	0	0		-10	5
8 09 Oct	-0.226881	70	-1	14		5	0	0		-12	4

Nafti Levante 23-24

Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia Posiciones	Racha	Despedido
0 14 Aug	-0.195786	113	-1	19		0	1	0		-1	0
1 22 Aug	-0.182218	244	-4	22		-3	1	0		-4	0
2 27 Aug	-0.040269	77	2	14		5	0	0		4	0
3 04 Sep	-0.162694	268	-1	18		1	0	0		0	3
4 11 Sep	-0.155010	272	-1	21		-2	1	0		-3	3
5 19 Sep	-0.159476	309	-2	21		-2	1	0		-3	3

Guede Malaga 22-23

	Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia	Posiciones	Racha	Despedido
0	15 Aug	-0.113043	76	-1	20		-1	1	0		-13	0	0
1	21 Aug	-0.019360	50	1	11		8	0	0		-4	0	0
2	28 Aug	-0.082720	46	1	8		11	0	0		-1	3	0
3	05 Sep	-0.050508	39	0	10		9	0	0		-3	6	0
4	10 Sep	-0.157306	31	0	10		9	0	0		-3	7	0
5	17 Sep	-0.180645	66	-1	14		5	0	0		-7	8	0
6	25 Sep	-0.220264	97	0	14		5	0	0		-7	5	0
7	02 Oct	-0.183857	187	-2	16		3	0	0		-9	3	0
8	09 Oct	-0.195565	51	0	16		3	0	0		-9	2	1

Bolo Real Oviedo 22-23

	Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia	Posiciones	Racha	Despedido
0	14 Aug	-0.146851	61	-1	18		1	0	0		-10	0	0
1	20 Aug	-0.182468	78	0	18		1	0	0		-10	0	0
2	26 Aug	-0.215326	103	-1	21		-2	1	0		-13	1	0
3	02 Sep	-0.161078	51	0	22		-3	1	0		-14	1	0
4	09 Sep	-0.175634	85	0	19		0	1	0		-11	2	0
5	18 Sep	-0.146819	306	-1	21		-2	1	0		-13	3	1

Cervera Real Oviedo 23

	Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia	Posiciones	Racha	Despedido
0	13 Aug	0.000000	3	-1	16		3	0	0		-7	0	0
1	18 Aug	-0.219550	6	-1	20		-1	1	0		-11	0	0
2	25 Aug	-0.172600	8	-1	22		-3	1	0		-13	0	0
3	02 Sep	-0.131767	6	1	17		2	0	0		-8	0	0
4	10 Sep	-0.083033	6	-2	20		-1	1	0		-11	3	0
5	16 Sep	-0.022769	13	-1	22		-3	1	0		-13	3	0
6	22 Sep	-0.017457	14	-1	22		-3	1	0		-13	3	1

del amo Cartagena 24

Y a continuación casos en los que el rendimiento del equipo era mediocre o bueno y no se planteó cambio de entrenador:

	Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia	Posiciones	Racha	Despedido
0	14 Aug	0.001475	20	4	1		18	0	0		15	0	0
1	20 Aug	-0.170043	7	-3	8		11	0	0		8	3	0
2	27 Aug	0.000000	2	-1	12		7	0	0		4	3	0
3	03 Sep	-0.042625	4	0	13		6	0	0		3	3	0
4	10 Sep	0.007567	12	1	8		11	0	0		8	4	0
5	17 Sep	-0.125833	3	-2	14		5	0	0		2	4	0
6	24 Sep	-0.008283	6	-2	19		0	1	0		-3	4	0
7	01 Oct	0.173711	9	1	13		6	0	0		3	4	0
8	05 Oct	-0.276450	6	-2	14		5	0	0		2	6	0
9	08 Oct	-0.170680	5	-1	17		2	0	0		-1	3	0

Alessio Mirandes 24

	Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia	Posiciones	Racha	Despedido
0	12 Aug	-0.106622	9	-4	22		-3	1	0		-17	0	0
1	19 Aug	0.000000	1	1	14		5	0	0		-9	0	0
2	28 Aug	-0.462100	1	-1	17		2	0	0		-12	3	0
3	03 Sep	-0.121914	7	-1	21		-2	1	0		-16	3	0
4	09 Sep	-0.184792	12	-1	22		-3	1	0		-17	3	0
5	17 Sep	-0.181975	4	2	17		2	0	0		-12	3	0
6	22 Sep	0.175075	4	1	12		7	0	0		-7	3	0
7	30 Sep	0.011000	4	3	9		10	0	0		-4	6	0
8	05 Oct	0.000000	4	2	8		11	0	0		-3	9	0
9	08 Oct	0.000000	2	1	7		12	0	0		-2	12	0

Joseba Eibar 24

	Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia	Posiciones	Racha	Despedido
0	11 Aug	-0.234050	28	0	13		6	0	0		5	0	0
1	19 Aug	-0.077294	18	1	5		14	0	1		13	1	0
2	25 Aug	-0.122758	12	1	4		15	0	1		14	4	0
3	02 Sep	-0.093329	17	0	6		13	0	1		12	7	0
4	08 Sep	-0.093361	31	-3	7		12	0	0		11	8	0
5	16 Sep	-0.118419	16	2	5		14	0	1		13	7	0
6	23 Sep	-0.128517	6	2	5		14	0	1		13	7	0
7	30 Sep	-0.077478	9	0	4		15	0	1		14	7	0
8	03 Oct	-0.160172	25	0	6		13	0	1		12	7	0
9	06 Oct	0.021705	22	2	4		15	0	1		14	8	0

Calleja Levante 24

Partido	Sentimiento Global	Numero de Tweets	Resultado	Posición	Distancia	Descenso	En Descenso	En Playoff	Diferencia	Posiciones	Racha	Despedido
0 13 Aug	-0.128609	11	0	9		10	0	0		-3	0	0
1 20 Aug	0.041225	8	0	15		4	0	0		-9	1	0
2 27 Aug	0.097900	4	-2	19		0	1	0		-13	2	0
3 01 Sep	0.183891	11	2	12		7	0	0		-6	2	0
4 09 Sep	0.075000	16	-2	15		4	0	0		-9	5	0
5 16 Sep	0.126988	8	1	12		7	0	0		-6	4	0
6 23 Sep	0.000000	3	-1	15		4	0	0		-9	6	0
7 30 Sep	-0.221975	4	2	11		8	0	0		-5	6	0
8 03 Oct	-0.148000	2	1	10		9	0	0		-4	6	0
9 06 Oct	-0.169213	15	-2	11		8	0	0		-5	9	0

Albes Albacete 24

Se podrían añadir más columnas teniendo en cuenta el número de aficionados del equipo. Al final clubes grandes como el Real Oviedo pueden ejercer una mayor presión y por ello que sean de los primeros equipos en tomar una decisión si la racha inicial no es buena.

3. Próximos objetivos

Este proyecto no se ha podido completar en su totalidad al ser muy ambicioso para las últimas dos semanas del curso junto a más tareas a realizar y estudiar para un examen final. Sin embargo, considero que tiene un potencial académico muy importante para tener un buen portfolio.

1. En primer lugar, realizar un sistema de puntuaciones con un decision tree algo casero para determinar el rendimiento del entrenador. Por supuesto para ello se necesitarán al menos 40 casos y solo dispongo de 10.
2. En segundo lugar, un estudio más profundo de la parte de LNP, para tener una mayor sensibilidad de la opinión pública.
3. En tercer lugar, aplicar series temporales para realizar un estudio de la evaluación del análisis de sentimientos, al final no es lo mismo ganar un partido que ganar el tercer partido consecutivo. Al igual que no es lo mismo empatar un partido a llevar diez partidos sin ganar.
4. Por último, una API con Django para poder evaluar la situación de tu equipo en tiempo real(esto sería un poco optimista porque necesitarías la API de Twitter).

4. Conclusiones

Profundizar en este proyecto me permitirá abrirme puertas en el mundo laboral en uno de los temas actualmente más demandados, el Procesamiento de Lenguaje Natural.