Análisis de Rumores de Fichajes en Twitter

Matías Torres, Juan Mamani, Benjamín Espinoza, Diego De La Sotta

28 de Noviembre 2024

Abstract

Este informe presenta el análisis de la difusión de rumores de fichajes en Twitter durante las ventanas de traspasos. El objetivo principal fue identificar los jugadores y equipos más mencionados, así como los patrones de propagación de los rumores en la red social. Se emplearon técnicas de análisis de sentimiento, modelado de tópicos y análisis de redes sociales para obtener una visión clara de la dinámica de los rumores. Los resultados incluyen la identificación de los principales influencers y las tendencias emocionales asociadas a los fichajes.

Contents

1 Introducción

1.1 Contexto y Motivación

El presente proyecto tiene como objetivo analizar los rumores de fichajes que circulan en Twitter durante las ventanas de traspasos. Estos rumores, ampliamente discutidos en las redes sociales, juegan un papel crucial en la percepción pública de jugadores y equipos. Los datos en inglés fueron seleccionados para simplificar el análisis lingüístico y abarcar el mercado global más amplio.

1.2 Propósito del Proyecto

El propósito del proyecto es identificar los jugadores más mencionados, analizar los patrones de propagación de los rumores y evaluar el sentimiento general en torno a los mismos. Para ello, se utilizaron técnicas avanzadas de análisis de datos y procesamiento de lenguaje natural.

1.3 Estructura del Informe

El informe se estructura de la siguiente manera:

- **Objetivos**: Se presentan los objetivos del proyecto.
- **Metodología**: Se describe la metodología utilizada para la recolección, preprocesamiento y análisis de los datos.
- **Resultados**: Se detallan los hallazgos principales obtenidos a partir del análisis.
- **Discusión**: Se analiza la interpretación de los resultados y se abordan las limitaciones del estudio.
- **Conclusiones**: Se resumen los resultados clave y se proponen recomendaciones.

2 Objetivos

2.1 Objetivo General

Analizar la difusión de rumores de fichajes en Twitter, enfocándose en los tweets en inglés para identificar los jugadores más mencionados y los patrones de propagación de información.

2.2 Objetivos Específicos

- Identificar los nombres de jugadores y equipos más mencionados en rumores de fichajes en inglés.
- Analizar la red de interacciones en Twitter para determinar los principales influencers en la propagación de rumores.
- Evaluar el sentimiento general en torno a los rumores de fichajes.
- Visualizar los resultados de manera clara y comprensible para identificar patrones y tendencias.

3 Justificación

Este proyecto se enfoca en los rumores de fichajes en inglés para reducir la complejidad del análisis lingüístico y concentrarse en el mercado global más amplio. La difusión de rumores en redes sociales como Twitter influye directamente en la percepción pública de jugadores y equipos, lo que puede tener un impacto significativo en las decisiones de los clubes, agentes y medios de comunicación.

4 Planificación del Proyecto

A continuación, se describen las fases del proyecto, los plazos estimados y los responsables de cada tarea.

4.1 Plan de Trabajo

El proyecto se estructuró en varias fases iterativas, las cuales permitieron un desarrollo continuo y la mejora de los modelos y prototipos. A continuación, se presenta el cronograma detallado con las fases y sus responsables:

Fase	**Descripción**	**Duración**	**Responsables**
Fase 1: Definición	Formulación del problema y definición de objetivos.	1 semana	Grupo completo
Fase 2: Recopilación	Configuración y uso de la API de Twitter para recolectar datos.	2 semanas	Matías Torres y Juan Mamani
Fase 3: Preprocesamiento	Limpieza y normalización de los datos recolectados.	1 semana	Benjamín Espinoza y Diego De La Sotta
Fase 4: Análisis	Aplicación de modelos de tópicos, análisis de redes y sentimiento.	2 semanas	Juan Mamani
Fase 5: Desarrollo	Creación del prototipo interactivo en Jupyter Notebook.	2 semanas	Diego De La Sotta
Fase 6: Iteración	Pruebas y refinamiento del prototipo.	1 semana	Matías Torres y Benjamín Espinoza
Fase 7: Documentación	Redacción del informe final y preparación de la presentación.	1 semana	Grupo completo
Fase 8: Presentación	Presentación oral y entrega del informe final.	1 día	Grupo completo

Table 1: Plan de trabajo y cronograma del proyecto

4.2 Metodología Iterativa

El enfoque adoptado para este proyecto fue iterativo, permitiendo ajustes y mejoras continuas en las distintas fases. Durante el desarrollo, las siguientes metodologías y herramientas fueron fundamentales:

- Recopilación de Datos: En un inicio, se utilizó la API de Twitter con bibliotecas como Tweepy y snscrape para recolectar los tweets relacionados con los rumores de fichajes en inglés. La recolección se centró principalmente en los términos específicos de fichajes, utilizando filtros para obtener una muestra representativa.

- Preprocesamiento de Datos: Se aplicaron técnicas de limpieza para eliminar URLs, menciones, hashtags irrelevantes y caracteres especiales. El preprocesamiento permitió organizar los datos de forma eficiente y enfocarse solo en los tweets relevantes.

- Análisis de Sentimiento: Se utilizó un modelo de sentimiento preentrenado, como RoBERTa, para clasificar el tono de los tweets (positivo, neutral, negativo). Además, se aplicaron técnicas de análisis de redes sociales utilizando NetworkX para identificar los principales jugadores y equipos mencionados y mapear las interacciones.

- Iteración de Modelos: Durante el proyecto, se realizaron iteraciones continuas en los modelos de análisis de sentimiento y redes sociales para mejorar la precisión y eficiencia de los resultados. Las pruebas continuas ayudaron a ajustar los parámetros y mejorar el rendimiento general del análisis.

4.3 Ajustes y Desafíos

A lo largo del proyecto, se enfrentaron ciertos desafíos que llevaron a ajustes en el cronograma original:

- Limitaciones de la API de Twitter: Debido a las restricciones de acceso de la API, se implementaron soluciones adicionales como el uso de snscrape para ampliar la recolección de datos, lo que permitió superar algunas limitaciones de la API de Twitter.

- Tiempo de Recolección de Datos: El proceso de recolección de datos fue más lento de lo esperado, lo que llevó a ajustar el cronograma y asignar más tiempo a la recolección de datos y el análisis iterativo de los modelos.

- Ajuste de Modelos de Sentimiento: Debido a la naturaleza informal y las abreviaciones en Twitter, se ajustaron los modelos de análisis de sentimiento para mejorar su precisión, especialmente al analizar tweets con ironía o abreviaciones.

5 Metodología

5.1 Recopilación de Datos

- **Herramientas**: Para la recolección de los datos, utilizamos la herramienta Playwright junto con cookies para simular la navegación en Twitter sin usar la API oficial. Playwright permite la automatización de navegadores y es más eficiente que las APIs tradicionales al interactuar

con la interfaz de usuario de un sitio web. Utilizamos cookies válidas obtenidas a partir de sesiones anteriores para garantizar el acceso a datos de Twitter sin limitaciones de las API. Además, implementamos scrolls dinámicos y pausas aleatorias para emular el comportamiento humano y evitar ser detectados como bots por el sistema de detección de Twitter.

- **Proceso**: El proceso implicó simular la navegación en la página de búsqueda de Twitter, desplazar el contenido dinámicamente para cargar más tweets y recoger datos relevantes sobre rumores de fichajes. Esto se hizo con configuraciones específicas de velocidad y pausas entre acciones, como un scroll aleatorio y pausas de entre 2 y 5 segundos, ajustando el comportamiento para que el proceso de recolección fuera lo más natural posible.

- **Muestra de Datos**: Para el análisis, se seleccionaron 10 tweets relevantes relacionados con rumores de fichajes, y se recolectaron hasta 200 respuestas por cada tweet para obtener una muestra representativa de las interacciones. Además, debido a la influencia de fuentes como Fabrizio Romano en la difusión de rumores, se priorizó la recolección de tweets de cuentas influyentes dentro del ámbito futbolístico.

5.2 Preprocesamiento de Datos

- Se realizaron técnicas de limpieza de datos, eliminando URLs, menciones, hashtags irrelevantes y caracteres especiales.

- Se filtraron los tweets para centrarse únicamente en aquellos que contenían rumores de fichajes de jugadores y equipos.

5.3 Análisis y Modelado

- **Análisis de Tópicos**: En lugar de utilizar modelos complejos como LDA (Latent Dirichlet Allocation), optamos por un enfoque basado en keywords. Esta metodología consiste en la búsqueda de términos clave relacionados con los rumores de fichajes. Se seleccionaron keywords amplias como "loan", "deal", "transfer", "rumor", "uncertain", "future", "signing". Este enfoque es flexible y puede ajustarse según el contexto, permitiendo el uso de nombres de jugadores o equipos específicos para focalizar el análisis en aspectos particulares. Debido a la alta influencia de figuras como Fabrizio Romano, se incluyó también su nombre como palabra clave en ciertos análisis, ya que sus publicaciones son fundamentales en la propagación de rumores en el ámbito futbolístico.

- **Análisis de Redes Sociales**: Utilizando la biblioteca NetworkX, se mapeó la red de interacciones en Twitter para detectar a los principales jugadores de los rumores. Esto incluyó no solo a los jugadores y equipos mencionados, sino también a las figuras influyentes en la difusión de estos rumores, como Fabrizio Romano.

- **Análisis de Sentimiento**: Para evaluar el sentimiento de los tweets relacionados con rumores de fichajes, se utilizó un modelo preentrenado basado en RoBERTa específicamente ajustado para Twitter: cardiffnlp/twitter-roberta-base-sentiment. Este modelo fue elegido debido a su capacidad para manejar el lenguaje natural y las abreviaciones comunes en las redes sociales. El modelo clasifica los tweets en tres categorías de sentimiento: positivo, neutral y negativo.

¿Qué hace este modelo?

RoBERTa es un modelo basado en BERT (Bidirectional Encoder Representations from Transformers), pero con algunas mejoras, como un mayor tamaño y entrenado con más datos. El modelo cardiffnlp/twitter-roberta-base-sentiment está ajustado específicamente para clasificar tweets, que suelen tener un lenguaje informal, abreviaciones y emojis, lo que lo hace ideal para este tipo de análisis.

¿Por qué se eligió este modelo?

Este modelo es más adecuado para el análisis de sentimientos en Twitter que otros modelos como PySentimiento, ya que está entrenado específicamente para entender el contexto y el tipo de lenguaje utilizado en esta red social.

5.4 Visualización de Resultados

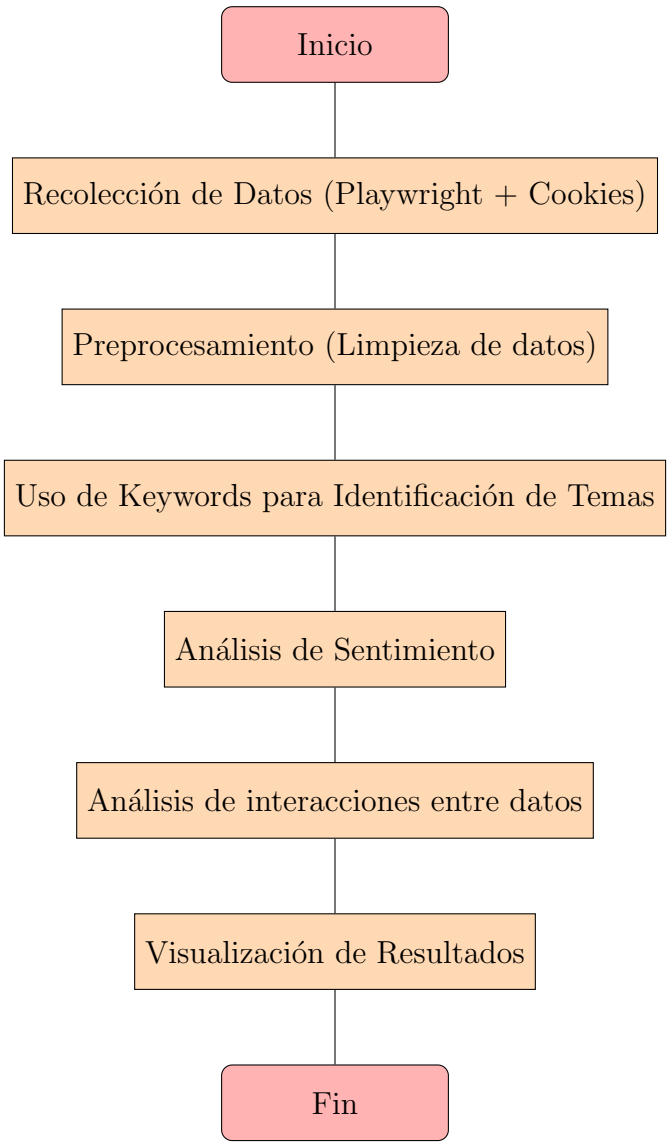
- Se crearon visualizaciones interactivas con matplotlib y seaborn para representar los hallazgos de manera clara.
- Se desarrolló un dashboard interactivo en Jupyter Notebook para facilitar la visualización y el análisis de los datos.

5.5 Desarrollo de Prototipo

El prototipo interactivo en Jupyter Notebook integró el análisis de datos y las visualizaciones. El prototipo fue iterado y refinado a lo largo del proyecto.

5.5.1 Diagrama del proceso del proyecto

A continuación, se presenta un diagrama que ilustra el flujo de trabajo utilizado para la recolección de datos, análisis y visualización de resultados. Este flujo resume el proceso descrito en la metodología.



6 Resultados

6.1 Hallazgos Principales

Los resultados obtenidos a partir del análisis de sentimiento incluyen los siguientes aspectos clave:

- **Distribución de Sentimientos**: Los tweets y respuestas fueron clasificados en tres categorías de sentimiento: Positivo, Neutral y Negativo.
- **Sentimiento Promedio por Jugador**: Además de identificar los jugadores más mencionados, se analizó el sentimiento promedio de los comentarios asociados a cada jugador, lo que permitió identificar a aquellos que generaron reacciones más polarizadas.

6.2 Análisis de Sentimiento

El análisis de sentimiento fue realizado utilizando un modelo preentrenado RoBERTa, específicamente ajustado para Twitter. Este modelo clasificó los tweets y respuestas en las categorías Positivo,

Neutral y Negativo, proporcionando una visión clara de las emociones asociadas a los rumores de fichajes.

Distribución de Sentimientos

A continuación, se presenta la distribución de los sentimientos procesados en los tweets y sus respuestas:

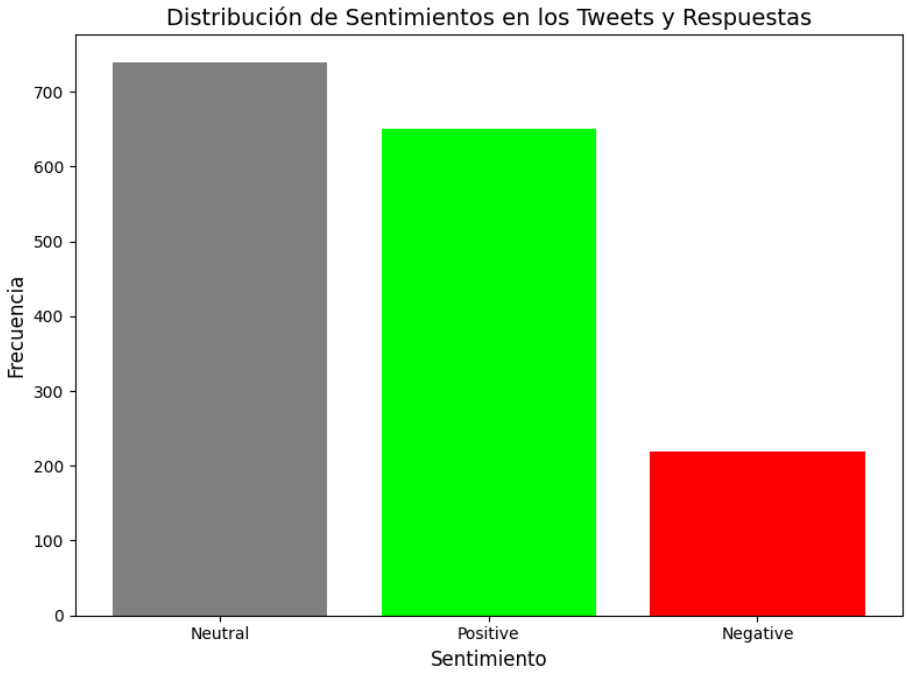


Figure 1: Distribución de los sentimientos en los tweets y respuestas.

Sentimiento Promedio

A través del análisis de las respuestas, se calculó el sentimiento promedio asociado con los jugadores y técnicos más mencionados en los rumores de fichajes. Los nombres que generaron mayor negatividad en las respuestas fueron Rodri, Pep Guardiola y Lionel Messi.

6.3 Gráficos y Wordcloud

Se presentarán a continuación los gráficos y nube de palabras que ilustran los hallazgos mencionados. Estas visualizaciones muestran la frecuencia de menciones y el sentimiento promedio.

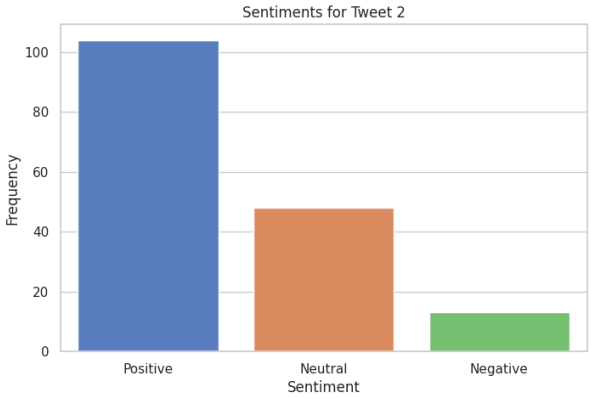


Figure 2: Tweet Principal 2: Paris Saint-Germain are closing in on the agreement with Nuno Mendes over new deal!

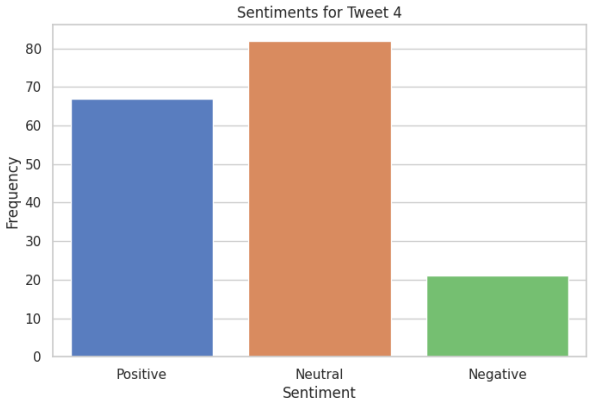
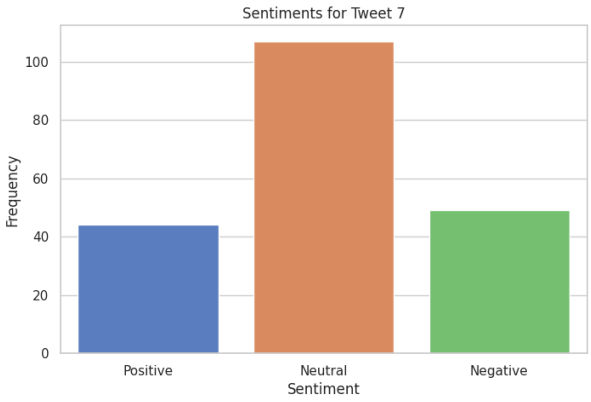


Figure 3: Tweet Principal 4: Pep Guardiola: "It took two hours to agree on new deal with Man City board".



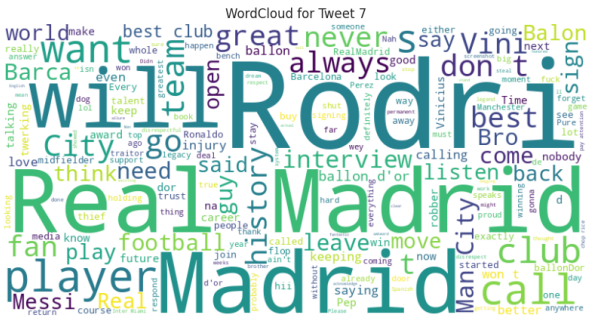


Figure 4: Tweet Principal 7: Rodri on potential future Real Madrid call: "When Real Madrid call you, the best club in history and the most successful... it makes you proud and you always have to listen".

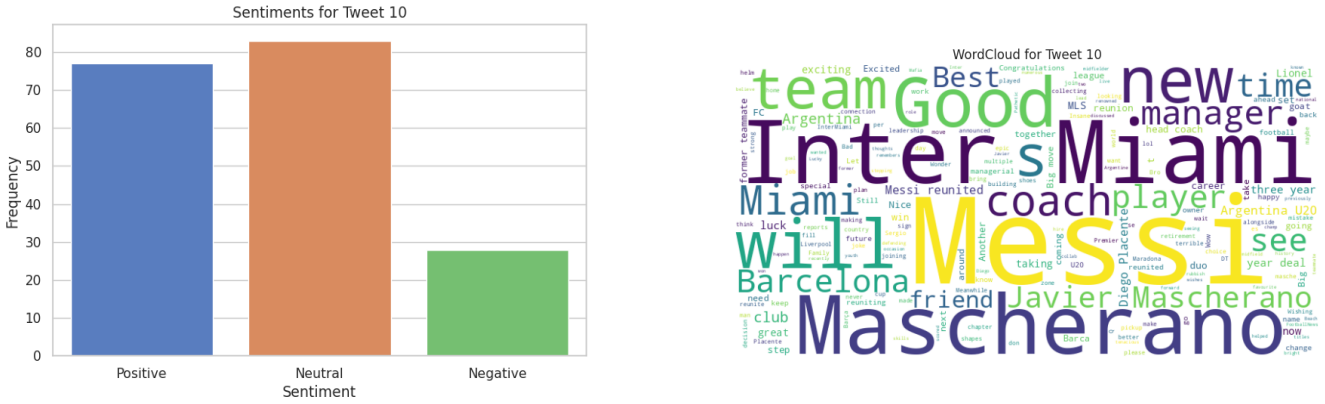


Figure 5: Tweet Principal 10: Javier Mascherano will be officially announced as new Inter Miami manager in the upcoming days.

6.4 Gráficos y Visualización de la Red de Interacciones

Para comprender mejor las relaciones entre los jugadores, equipos mencionados en los rumores de fichajes, se construyó un grafo de interacciones que muestra cómo los jugadores están conectados a través de menciones conjuntas. Cada nodo representa un jugador, y las aristas entre ellos indican que fueron mencionados juntos en los mismos tweets o respuestas.

A continuación, se presenta una visualización estática del grafo, donde:

- El tamaño de los nodos refleja la frecuencia de menciones de cada jugador.
- El color de los nodos refleja el sentimiento promedio hacia el jugador: verde para sentimiento positivo, rojo para sentimiento negativo y gris para sentimiento neutral.

La siguiente figura muestra el grafo de interacciones entre los jugadores basado en menciones conjuntas y sentimiento promedio:

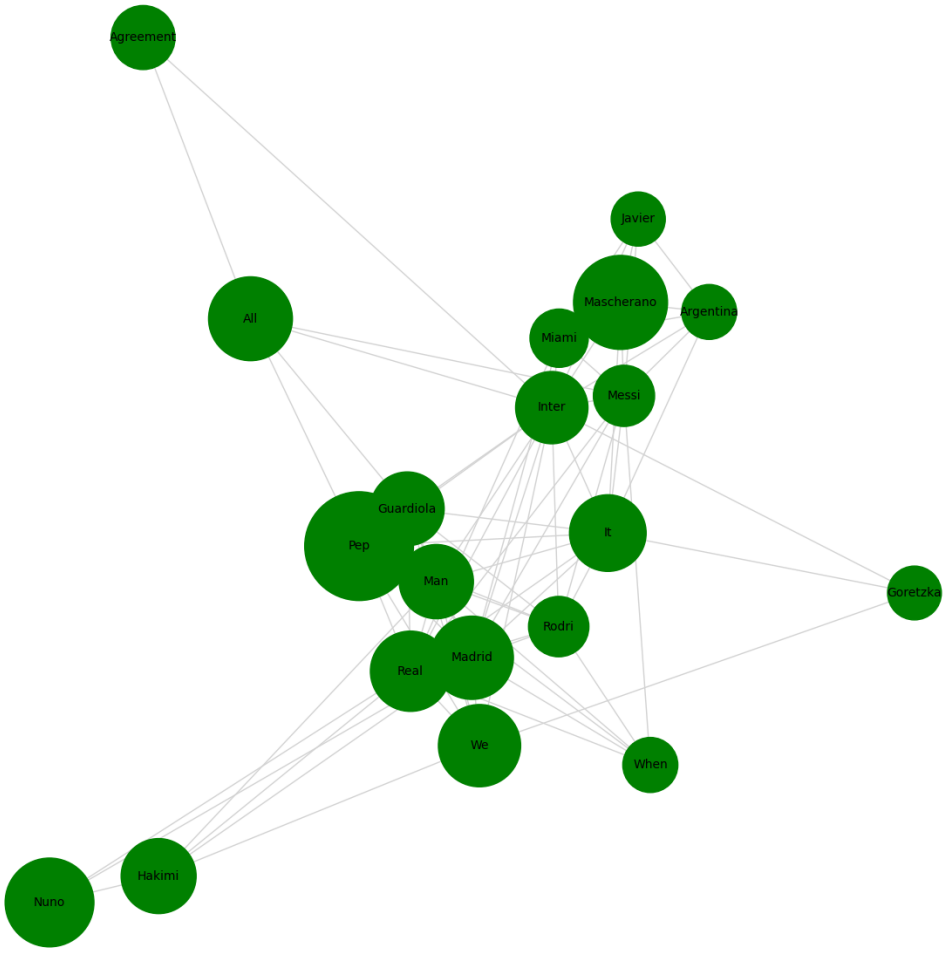


Figure 6: Grafo de interacciones basado en menciones conjuntas y sentimiento promedio.

7 Discusión

7.1 Interpretación de los Resultados

Los resultados muestran que los rumores de fichajes se difunden de manera significativa a través de las redes sociales, con ciertas figuras clave desempeñando un papel importante en la propagación de estos rumores. El análisis de sentimiento revela que la opinión pública es ambigua en cuanto a muchos fichajes, lo que puede influir en la toma de decisiones de clubes y jugadores. En particular, las figuras influyentes como Fabrizio Romano tienen un impacto notable en la difusión y percepción de estos rumores, ya que sus publicaciones en Twitter generan un gran volumen de interacción y son clave en la propagación de la información.

7.2 Limitaciones

El proyecto enfrentó las siguientes limitaciones:

- **Recopilación de Datos**: La API de Twitter impuso restricciones, por lo que utilizamos Playwright + cookies para obtener más datos. Sin embargo, esto hizo que la recolección fuera más lenta y dependiera de la estabilidad de la página. Además, debido a la alta influencia de figuras como Fabrizio Romano, nos centramos principalmente en sus tweets, lo que limitó la diversidad de las fuentes de datos.

- **Muestra Limitada**: Solo se recopilaron 10 tweets y 200 respuestas por tweet, lo que dificultó el análisis de los autores de los rumores. Esto nos llevó a enfocarnos más en el análisis de sentimientos hacia los jugadores, sin explorar a fondo la influencia de otras fuentes.

- **Sesgo de Datos**: El uso de keywords amplias introdujo cierto sesgo, ya que algunos tweets no eran completamente relevantes para el tema de fichajes. Este sesgo se incrementó al centrarse principalmente en los tweets de Fabrizio Romano, lo que restringió el análisis a una sola fuente influyente en el ámbito futbolístico.

8 Conclusiones

El proyecto alcanzó los objetivos principales, con énfasis en la identificación de los jugadores más mencionados y el análisis de sentimiento hacia los rumores de fichajes. Los resultados mostraron una distribución mixta de sentimientos (positivo, neutral y negativo). Sin embargo, debido a las limitaciones en la recolección de datos, no se pudo profundizar en el análisis de los autores de los rumores, lo cual fue una de las principales limitaciones. El análisis se centró especialmente en Fabrizio Romano, una figura influyente que desempeña un papel crucial en la propagación de rumores de fichajes.

Se utilizó Playwright + cookies para superar las restricciones de la API, lo que permitió acceder a más datos. Sin embargo, la muestra obtenida (10 tweets y 200 respuestas por tweet) fue limitada, lo que restringió el alcance del análisis.

8.1 Cumplimiento de Objetivos

- Objetivo 1: Identificar los jugadores más mencionados. Cumplido. Se identificaron los jugadores más recurrentes en los rumores de fichajes.

- Objetivo 2: Analizar los patrones de propagación de los rumores. Parcialmente cumplido. El análisis se centró en los jugadores, ya que los datos sobre los autores de los rumores fueron limitados. La influencia de Fabrizio Romano fue clave en la propagación de ciertos rumores.

- Objetivo 3: Evaluar el sentimiento general. Cumplido. Se realizó un análisis completo de sentimiento, utilizando el modelo RoBERTa.

- Objetivo 4: Visualizar los resultados. Parcialmente cumplido. Las visualizaciones de sentimiento y menciones fueron realizadas, pero el análisis de los autores de los rumores no fue posible debido a la falta de datos.

8.2 Recomendaciones

Se recomienda mejorar la calidad y cantidad de los datos recolectados, ampliando la muestra para incluir más tweets y respuestas de otras fuentes influyentes, como Fabrizio Romano. Además, sería

útil refinar el modelo de análisis de sentimiento para abordar mejor las características del lenguaje informal en Twitter, y diversificar las fuentes de datos para capturar una gama más amplia de opiniones sobre los rumores de fichajes.