Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Analizando el Campeonato de Primera División 2016-17 de Argentina en la red social Twitter

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Brian Litwak

Director: Lic. Ernesto Mislej

Buenos Aires, 2017

ANALIZANDO EL CAMPEONATO DE PRIMERA DIVISIÓN 2016-17 DE ARGENTINA EN LA RED SOCIAL TWITTER

¿Cómo afecta el resultado del futbol local en el humor de los usuarios en Twitter? ¿Es posible inferir el estado de ánimo de los fanáticos en función de los resultados de sus equipos? ¿Cuánto dura la alegría y cuánto la tristeza? ¿Y en los clásicos? ¿La hinchada del equipo ganador suele estar más confiada antes del comienzo del partido? ¿Es posible inferir el resultado del partido sondeando la previa?

Desarrollamos una base de datos de perfiles de Twitter y propusimos un procedimiento para inducir el club el cual son hinchas. Además, implementamos una herramienta que obtiene la información de Twitter de forma eficiente.

Realizamos un análisis de sentimiento a los tweets mediante 3 métricas creadas por nosotros. Y finalmente entrenamos modelos para predecir el resultado del partido, la cantidad de goles marcados, y la diferencia de gol utilizando información obtenida de la red social.

Palabras claves: Twitter, Análisis de sentimiento, predecir resultados de partidos, clasificar usuarios,Torneo de primera división de argentina.

ANALYZING THE 2016 - 17 ARGENTINE PRIMERA DIVISION IN THE SOCIAL NETWORK TWITTER

This thesis work will investigate if it is posible to predict the mood of football fans in Twitter, based on match results of their respective teams, and if it is possible to predict a match result based the mood of fans in Twitter. In order to achieve that, we studied the behavior of Argentina football fans in the social network Twitter with the aim of building a predictor of match results. First, we build a database of Twitter users labeled with their favorite club. Then, we proposed a procedure to find out which fans belong to each team, and used a tool to optimize data collection from Twitter. After that, we performed a sentimental analysis of tweets by using our own metrics criteria. Finally, we trained models to predict match's resuts, goals scored and the goal difference using team's stats and information from Twitter, which gives an enhanced score. We conclude that it is possible to infer the mood of football fans in Twitter and our best model obtains an accuracy of 0.63 % when modelling match results.

Keywords: Twitter, sentiment analysis, predict match results, classify users, First Division Tournament of Argentina.

A mis viejos por permitirme dedicarme a estudiar en toda mi travesía facultativa, por bancarme cuando volvía de mal humor o quemado, y esperarme siempre para cenar. Estas dos líneas no alcanzan para agradecerles por todo lo que hicieron y hacen por mi.

A Ernesto por aceptarme como tesista, proponerme un tema que me encanta, y soportar millones de whatsapp míos sobre la tesis.

A Mati Alvarez por darme siempre una mano con la facu, ser el mejor compañero de trabajos prácticos que se puede tener, siendo no solo un compañero de lujo sino también un gran amigo.

A mis amigos de Hangover y Bria porque a pesar de no entenderme muchas veces lo mucho que estudió produciendo mi ausencia en algunas salidas, siempre están ahí.

A Maia por siempre guiarme en mi carrera académica y profesional y darme varias manos en la elaboración de esta tesis.

A la UBA por permitirme estudiar lo que me gusta con docentes de calidad que lo hacen por pasión y gusto.

A la gente de MeaningCloud por brindarme una cuenta gratuita para poder usar su herramienta en mi tesis.

A todos los que me dieron una mano con la tesis.

HALT.

Índice general

1. INTRODUCCIÓN

Twitter ha sido fundada en el 2006. En el transcurso de estos 11 años, se convirtió en uno de las redes sociales más importantes con una base de 328 millones de usuarios activos por mes que producen 500 millones de posteos en la red social llamados Tweets por día.

Tal ha sido usado para muchos trabajos científicos de diverso tipo como por ejemplo clasificar usuarios por su edad, sexo, ubicación, nacionalidad ,orientación política o religiosa y otros intereses [1] [2] [3] [4] .También para detectar enfermedades[5] ,resumir eventos [6] [7] ,clasificar texto [8] [9] y análisis de sentimiento.

Sin embargo no encontramos un estudio del futbol de Argentina en Twitter. Encontramos otras publicaciones similares pero no eran en la Argentina o utilizaban otro deporte.

Por ejemplo, una investigación de Facebook [1] o trabajos científicos utilizando Twitter sobre la NFL [10]

También encontramos publicaciones similares utilizando como deporte al fútbol y en la liga de Inglaterra donde en una tampoco realizan análisis de sentimiento [11] y otra publicación de IBM donde hacen análisis de sentimiento del minuto a minuto de la definición de la Premier League [2]

En el mismo país, bajo los juegos olímpicos de Londres 2012 se realizó un análisis de sentimiento en Twitter, y en base a si los tweets de ese dia fueron en su mayoría positivos, neutral, o negativos, la atracción turística de London Eye cambiaba de color[3]. Este evento mundial deportivo fue el primero en contener un espectáculo de luces utilizando análisis de sentimiento en las redes sociales.

Optamos por realizar el estudio con datos de la Argentina no solo porque es novedoso sino por los siguientes dos motivos.

- Por un lado, Twitter en Argentina tiene alrededor de 11 millones de usuarios con una penetración del 70 por ciento en el mercado y por ejemplo, para las últimas elecciones presidenciales durante las principales jornadas relacionadas al tema - los dos debates, la primera vuelta y el ballottage - se efectuaron 11 millones de tuits desde la plataforma [4]

- Por otro lado el fútbol es el principal deporte en la Argentina moviendo multitudes a lo largo y ancho del país generando un montón de comentarios en las redes sociales. Para tener una idea de lo que mueve esta pasión en Argentina,el Havas Group realizó un informe del Torneo de la Independencia en la televisación donde señala que el Share máximo alcanzado en el Superclásico del 2016 fue del 58 %

[1] investigación de Facebook
[2] Publicación de IBM
[3] Noticia en telegraph
[4] Entrevista al vicepresidente para América latina de Twitter

entre los televidentes argentinos expuestos. Además, un 84 % de los argentinos vio al menos un minuto de alguno de los encuentro [5].

Un hito de Twitter es que la final de la copa mundial de 2014 disputada en Brasil donde la Argentina llegó a tal instancia obtuvo la mayor cantidad de los tweets por minuto siendo 618.725 tweets por minuto demostrando no solo el interés por el fútbol de los Argentinos sino a nivel mundial [6].

El Campeonato de Primera División 2016/2017 de la Argentina, también conocido como Torneo de la Independencia, comenzó el 26 de Agosto de 2016 finalizando el 27 de junio de 2017 donde se disputaron 30 fechas y participaron 30 clubes. Se disputó a una sola rueda, por el sistema de todos contra todos, con el agregado de una fecha especial de clásicos [7].

Sumando los partidos internacionales y los disputados por copa Argentina y restando el receso, se disputó un partido por semana donde fueron transmitidos por televisión abierta para los Argentinos. Esto abre la oportunidad que los fanáticos del fútbol, puedan expresar en las redes sociales no solo su opinión del equipo por cual simpatizan, sino de cualquier equipo.

[5] Noticia sobre el informe del Havas Group
[6] Noticia sobre el récord producido en la final de la copa mundial de 2014 de fútbol
[7] Reglamento del campeonato de primera división 2016/2017

1.0.1. Objetivo

Lo que proponemos es tratar de resolver las siguientes dos preguntas:

- ¿Se puede predecir el estado de ánimo en la red social Twitter de los hinchas de fútbol en base a los resultados de los partidos de fútbol de sus respectivos equipos?
- ¿Se puede pronosticar el resultado de un partido de fútbol en base al estado de ánimo en la red social Twitter de los hinchas de fútbol de ese equipo?

Ambas preguntas las analizamos sobre un grupo de equipos de futbol de primera división del fútbol argentino de la temporada 2016/2017 y definimos el estado de ánimo de sus hinchas basándonos en los tweets escritos.

Notar que la segunda pregunta es estudiada a nivel mundial donde no solo existen muchas publicaciones científicas en Google Scholar [8], alrededor de 9100 resultados, sino también a nivel industrial ya que existen empresas que se dedican a construir modelos para predecir resultados de deportes. [9].

[8] Búsqueda en Google Scholar de predecir resultados de partidos de fútbol
[9] Algunos empresas que se dedican a esto son Stratagem, Mustards,y SmartOdds

1.0.2. Estructura de la tesis

En el capítulo 2 se presentan distintas técnicas que se utilizan actualmente para resolver tres problemas que nos interesan en esta tesis, los cuales son:

- Clasificar a perfiles en Twitter
- Realizar análisis de sentimiento a fragmentos de texto
- Predecir resultados deportivos

En el capítulo 3 se presenta como nosotros encaramos los primeros dos problemas mientras que el último y tercer problema lo encaremos en el capítulo 6. Por un lado en el capítulo 4 describimos cuestiones de la experimentación como el dominio de datos y las métricas utilizadas. Por otro lado en el siguiente capítulo presentamos los resultados de esta experimentación. En el capítulo 7 se encuentran nuestras conclusiones de la investigación y en el capítulo 8 hay posibles ideas para continuar el trabajo. Por último en los capítulos 9 y 10 se encuentran unos anexos sobre más resultados de la experimentación del comportamiento del hincha Argentino y de desarrollo respectivamente

1.0.3. Motivación

Los motivos por los cuales nos parece esta tesis útil son

- Al momento de realizar esta tesis, no existe un estudio similar que trate de conocer al hincha Argentino en la red social Twitter
- Poder armar un predictor para inferir resultados del fútbol Argentino
- Nos pareció una buena forma para que la gente se pueda interesar tanto en la investigación como en la carrera a través del fútbol que es una pasión que mueve multitudes en Argentina

2. ESTADO DEL ARTE

Buscamos trabajos de investigación y publicaciones similares a los objetivos de nuestra tesis.

- Clasificar a usuarios en Twitter. En caso de clasificarlos por equipos de futbol mejor
- Analizar el sentimiento de Tweets sobre partidos y campañas de equipos de fútbol
- Predecir el resultado del partido de fútbol en base a información previa

2.1. Clasificar a usuarios en Twitter

2.1.1. Clasificar en base al equipo por cual es fanático

2.1.1.1. Using Twitter to predict football outcomes

En el paper [11] en realidad clasifican a los Tweets en lugar de sus perfiles en Twitter. La idea que usan es **armar una lista de hashtag para cada equipo y en caso de que un hashtag figure en un Tweet, ese Tweet pertenece al mismo equipo que pertenece ese hashtag**. En caso de que existan más de un hashtags en el Tweet que pertenecen a diferentes equipos, no se lo etiqueta con ningún equipo. La lista de hashtag se la armó con las cuentas oficiales de los equipos y con los apodos que pueden tener, pero **excluyendo los apodos que pueden ser de otro equipo de otros países** [1] Comentan en el paper que lo malo de esta técnica es que no le queda uniforme o similar la cantidad de Tweets por equipo ya que los equipos más populares de Inglaterra tienen mucho más Tweets que el resto de los equipos.
En el trabajo de investigación *Predicting the NFL Using Twitter* [10] utilizaron la misma metodología que en el explicado anteriormente pero para la NFL.

2.1.1.2. Classifying Twitter User Interests using Time Series

En la publicación [12] parten de la hipótesis que **los usuarios de Twitter que comparten un mismo interés, tienen también un patrón de periodicidad a la hora de twittear al mostrar su opinión o compartir su experiencia**. [2] Un usuario de Twitter va a ser representado por un conjunto de números que representan para cada ventana de tiempo y clase, la frecuencia en la que se utilizan palabras relacionadas de esa clase en esa ventana de tiempo

2.1.2. Clasificar por datos demográficos

Al momento de realizar esta tesis, no encontramos más papers que clasifiquen a los perfiles de Twitter por el equipo que es hincha en su deporte favorito pero la clasificación en Twitter está muy estudiada en los atributos de la edad, sexo, nacionalidad. Esto se debe a que Twitter es un buen lugar donde los usuarios se expresan, comparten información de una forma legítima dando lugar a que encuestadores y campañas de publicidad puedan usar la opinión de los usuarios pero para estos es muy importante poder segmentar a su encuestados por atributos demográficos. Por esto último esta muy estudiado el tema de la etiquetación de las cuentas de las red social en cuestión por los atributos antes mencionados.

[1] Por ejemplo para el equipo de fútbol de la liga inglesa Tottenham Hotspurs comparte el apodo de Spurs con el equipo de Basket de la NBA, San Antonio Spurs. Por ende, no usan como posible hashtag #Spurs.

[2] Por ejemplo, los fanáticos del deporte tienen más actividad durante el fin de semana que durante la semana ya que los partidos transcurren durante el fin de semana

Como al fin y al cabo, la primera parte es un problema de clasificación nos sirve también ver cómo se resuelve el problema de clasificar una cuenta de Twitter por su edad, sexo o nacionalidad.

2.1.2.1. Where's @wally? A Classification Approach to Geolocating Users Based on their Social Ties

En el paper [1] arma una base de perfiles de Twitter que cada perfil está clasificado por su ubicación y luego un modelo para dado un perfil sin etiquetar, obtener su ubicación.
Para armar el dataset primero usó un método computacional compuesto por reglas de experto sobre los perfiles y luego dos colaboradores verificaron 1000 perfiles del dataset generado obteniendo un 97.2 % de precisión al etiquetar con las técnicas computacionales. [3]
Lo que propone es contar la cantidad de seguidores y seguidos que se tiene en cada ciudad, y pertenece a la ciudad con mayor valor. La contra que tiene es que no es perfecto porque si todos tus amigos se fueron a vivir a otra ciudad a estudiar, te va a decir equivocadamente tu ciudad pero son los casos menores. Por ende propone poner peso a las ciudades. Es decir, un amigo de Charata te da más valor que un amigo de CABA. Utilizando la idea de entropía, los amigos de CABA otorgan poco valor de información mientras que un amigo de Charata te da más información. Si tenes un porcentaje importante de tus amigos de Charata, lo más probable es que seas de tal lugar.

2.1.2.2. An Interactive Method for Inferring Demographic Attributes in Twitter

En el paper [2] también clasifican los perfiles de Twitter por su sexo y edad. Separa en dos enfoques:

1. El enfoque basado en el perfil **utiliza metadata asociada a la cuenta del usuario de Twitter** [4]

2. El enfoque basado en el contenido del perfil explota **la utilización del lenguaje en los tweets del usuario**, generando n gramas como atributos para un algoritmo de aprendizaje, como por ejemplo un SVM [5]

[3] Estas reglas consisten en realizar primero un filtro a usuarios localizados en Inglaterra, que escriben en inglés, y que están dentro del uso horario GMT y analizar su descripción en Twitter en búsqueda de su ubicación

[4] Un ejemplo es asignar el sexo en base a buscar en un diccionario el nombre del perfil ya que muchos utilizan su verdadero nombre en las redes sociales. Es decir si en mi perfil figura Nicolás, lo busca en un diccionario de nombres y al encontrarlo devuelve sexo masculino para ese perfil

[5] Para poder etiquetar la edad del perfil de Twitter, se suele combinar ambos enfoques

Para generar el dataset, buscaron todos los tweets que un usuario de Twitter se autodesea un feliz cumpleanos con la edad, y se asigna a ese perfil la edad sugerida en el tweet extraída con expresiones regulares.

2.1.2.3. Inferring Latent Attributes of an Indian Twitter User using Celebrities and Class Influencers

En el trabajo de investigación [3] clasifica al usuario por su sexo, edad y afiliación política con perfiles de usuarios de India.
Su enfoque es predecir estos resultados **observando el comportamiento y el contenido lingüístico del perfil de Twitter** como el de las celebridades seguidas por el usuario. Para eso utiliza **la idea de influenciadores de clase**, que son perfiles de Twitter que influencian a una clase particular. Por ende, pueden discriminar a los de esa clase. Llaman celebridades a usuarios con mas de 10000 seguidores o las cuentas que hayan sido verificadas por Twitter y utiliza a las celebridades como influenciadores de clase [6]

- Los atributos de comportamiento son por ejemplo frecuencia de twitteo, tamaño promedio de tweet, dividir cantidad de seguidores con la cantidad de seguidos, frecuencia de hashtag, frecuencia de retweet, cantidad de celebridades seguidas, cantidad de links en tweets, cantidad de fotos en tweets, cantidad de emoticones promedio por tweet
- Los atributos lingüísticos son por ejemplo la cantidad de pronombres, conjunciones, y tipo de palabras (como por ejemplo negocios, deportes, moda)
- Los atributos del vecindario de famosos son por ejemplo la cantidad de famosos seguidos por cada edad y lo mismo del sexo. Para buscar el sexo de los famosos cuenta en wikipedia la cantidad de palabra ella o él

Siguiendo esta misma idea, el paper *Classifying Latent User Attributes in Twitter* [4] utiliza las palabras e iconos más frecuentemente usadas para ver el atributo de la edad y el sexo.

[6] Por ejemplo, si seguís a @laliespos o @TiniStoessel lo más probable que tu sexo sea femenino y seas una adolescente mientras que si seguís a @alfanograce tu sexo sea masculino y seas más adulto

2.2. Análisis de sentimiento de Tweets

Análisis de sentimiento es el proceso de determinar el tono emocional que hay detrás de una serie de palabras, y se utiliza para intentar entender las opiniones y emociones expresadas en un texto.

2.2.1. ¿Cómo funcionan las herramientas de análisis de sentimiento?

Aunque no nos proponemos realizar una herramienta de análisis de sentimiento de Tweets, nos informamos un poco de cuales son las ideas detrás de estas ya que vamos a utilizar ésta herramienta en el análisis de Tweets de los hinchas.

2.2.1.1. MeaningCloud

En la tesis utilizaremos el servicio de meaningcloud, por ende nos pareció correcto ver como funcionaba tal servicio [7]. Ésta utiliza enfoques semánticos, basados en un tratamiento avanzado del lenguaje natural en todos los aspectos de morfología, sintaxis, semántica y pragmática. El motor primero genera el árbol sintáctico-semántico del texto, y, sobre éste, aplica los términos del lexicon propagando los valores de polaridad a lo largo del árbol, combinando los valores de forma apropiada en función de la categoría morfológica de la palabra y de las relaciones sintácticas que los afectan teniendo en cuenta las posibles **configuración del usuario, como por ejemplo agregar una polaridad a un término para hacerlo más específico del tópico en cuestión**.

2.2.1.2. Trabajo de fin de grado de Lola Lage Garcia

Pero también existe otro enfoque donde dado un conjunto de Tweets etiquetados, se extraen atributos y se genera un modelo. Este es el caso del trabajo de fin de grado de Lola Lage Garcia[8] que utilizó 68.017 tweets que están etiquetados por su polaridad como corpus. Como herramientas utilizo Twitter4j que es una librería que se comunica con la api de Twitter para acceder a Tweets. También uso Freeling para dividir el texto en palabras o tokens, obtener su lema y etiquetar cada una de ellas con su correspondiente categoría gramatical y por último Weka para entrenar el clasificador, utilizando un modelo SVM. Antes de realizar un estudio del tweet, reemplazó las abreviaciones ya que en la red social se pueden utilizar 140 caracteres en cada estado. Por ejemplo, reemplaza salu2 por saludos. Además utiliza un fichero léxico de polaridad, lista de interjecciones con polaridad, y una lista de emoticones con polaridad que las utiliza para formar los atributos. Alguno de estos son:

- Número de interjecciones positivas

[7] Solución propuesta por MeaningCloud
[8] Trabajo de fin de grado de Lola Lage Garcia

- Número de interjecciones negativas
- Número de emoticonos positivos
- Número de emoticonos negativos
- Número de palabras positivas
- Número de palabras negativas
- Número de sustantivos
- Numero de adjetivos
- Número de adverbios
- Número de verbos
- Número de interjecciones

2.2.2. Otros estudios que utilizan análisis de sentimiento en el deporte

Cómo no vamos a implementar una herramienta de análisis, sino que la vamos a utilizar nos pareció interesante buscar otros estudios sobre análisis de sentimiento de Tweets en referencia al deporte.

2.2.2.1. Investigación de IBM sobre el desenlace de la Premier League 2011/2012

IBM realizó un análisis de sentimiento minuto a minuto en Twitter de la definición tanto del campeón de la Premier League 2011/2012 como qué equipo se mantiene en la Liga. Por el título tenían posibilidades el tanto el Manchester City como el Manchester United mientras que los que peleaban por no descender eran el QPR, que se enfrentaba al Manchester City y el Bolton. Con giros en el marcador, se puede ver como los fans de los equipos de Manchester reaccionan en Twitter ante un apasionante final [9].

2.2.2.2. Los juegos olímpicos 2012 en Londres

Estos fueron llamados los primeros juegos olímpicos en las redes sociales. Un hecho para ganar esa distinción fue que se realizaba un análisis de sentimiento a los tweets de usuarios de Inglaterra que hacían referencias a las olimpiadas cada noche, y en base al resultado se prendian las luces del London Eye a las 9 de la noche, convirtiéndose en el primer espectáculo mundial de luces dirigidas por redes sociales.[10]. En el paper *Social Media Analysis, Twitter and the London Olympics 2012* [13] analizan información de las redes sociales del mismo evento mundial llegando a los siguientes resultados

[9] Noticia sobre la publicación de IBM
[10] Noticia en telegraph sobre el espectáculo del London Eye al utilizar redes sociales

- La popularidad de los hashtag durante un periodo puede identificar eventos mundiales [11]
- Se puede ver como atributos demográficos pueden ser de interés [12]

Este evento mundial fue el comienzo de tratar a los medios de comunicación digital como una medida científica social.

2.2.2.3. Investigación de Facebook sobre la temporada 2013 de la NFL

La red social realizó un análisis de sentimiento de la temporada 2013 de la NFL [13] con el objetivo de ver cuánto al hincha le importa su equipo deportivo que deja que la performance de este afecte las emociones fuertemente.

Para medir el humor de los hinchas se utiliza una métrica donde la idea clave es **contar la cantidad de palabras positivas y negativas existen en cada estado**.

Se obtuvieron las siguientes conclusiones:

- Los hinchas tienen mejor humor en partidos ganados que en partidos perdidos
- Los fans de los equipos que eventualmente ganan tienden a ser un poco más positivos antes del partido que los del equipo que pierde [14]

[11] Tal es el caso del super Sabado donde Inglaterra ganó 3 medallas de oro que se destaca por la gran cantidad de Tweets que se registra en ese día a diferencia del resto de los días

[12] Por ejemplo las mujeres tienen un pico positivo mayor a los hombres mientras que estos últimos tienen picos negativos más grandes que ellas

[13] Análisis de sentimiento de la temporada 2013 de la NFL hecho por Facebook

[14] Esta última idea se concretó en 9 de 10 partidos de playoff mientras que durante la temporada regular se midió un 0.65 de AUC indicando que generalmente no es útil

2.3. Predecir el resultado del partido de fútbol

Twitter ha sido usado para predecir o explicar una variedad de eventos mundiales como por ejemplo el resultado de elecciones [14], que tan taquillera es una película antes de su estreno [15]
Trataremos de predecir los resultados de los partidos de fútbol con la información que tenemos de Twitter. Esta última idea no es muy novedosa ya que existen varios papers que cuentan su experiencia al tratar de predecir resultados de partidos de varios deportes utilizando las redes sociales como una de sus fuente de atributos. Sin embargo, nos sirvieron para reproducir la idea en este contexto para nuestro dominio de datos. También existen competencias a nivel mundial donde tratan de predecir los resultados de los partidos [15].

2.3.1. Using Twitter to predict football outcome

En este paper [11] que ya mencionamos anteriormente analizan si en base a datos estadísticos y Tweets se puede predecir resultados de la Premier League. Utilizando solamente a los primeros no dieron buenos resultados y demoraron mucho tiempo en conseguirse. **Sin embargo el mejor resultado se obtiene combinando ambos**.

Para armar los atributos basados en Tweets usa unigramas y bigramas pero filtrando por equipo y tipo de localia. Es decir se tienen distintos atributos para el equipo de local y visitante, como para cada equipo. Por ejemplo, San Lorenzo tienen unos atributos como local y otros como visitante que pueden ser iguales o no a los de Huracan. Entrenaron modelos Naïve Bayes, Random forests, Logistic regression y SVM.

El resultado que obtuvieron es que **utilizar entre 11 y 15 bigramas funcionan mejor que unigrama, y el clasificador que mejor resultado obtuvo fue el Random Forest para la combinación de ambos tipos de atributos**.

2.3.2. Predicting the NFL Using Twitter

En la publicación [10] que ya mencionamos antes tratan de predecir los resultados de partidos de la NFL en base a datos estadísticos como a Tweets.
Algunos de los atributos estadísticos:

- Porcentaje de victorias del equipo en condición de local para el local en la temporada actual mientras que el porcentaje de victorias del equipo visitante en condición de visitante en el mismo torneo.
- Promedio de puntos anotados
- Promedio de la diferencia entre puntos anotados y puntos recibidos

[15] Un ejemplos es March Machine Learning Mania 2016 en kaggle

Un dato de color es que **para atributos obtenidos de la red social utiliza la cantidad de veces que se habla del árbitro** ya que encontraron una relación que si se habla mucho del árbitro y el equipo es visitante, existe una posibilidad mayor a que el equipo pierda. También la **cantidad de veces que figura en los tweets palabras como genial o ganar**.
Llegan a la conclusión que **utilizar modelos basados en atributos derivados de la red social tienen mejor performance que modelos que sólo utilizan atributos de las estadísticas**.

2.3.3. TwitterPaul: Extracting and Aggregating Twitter Predictions

Por último en el trabajo de investigación [16] tratan de predecir los resultados de los partidos del mundial de fútbol en sudáfrica en el 2010.

Para eso separan la extracción de la predicción en los tweets de la predicción del resultado del partido. Para hacer la **extracción de la predicción en un tweet usan una gramática libre de contexto para parsear el Tweet** mientras que para las predicciones propone dos formas

1. Contar la cantidad de veces que aparece el nombre del equipo en los tweets en los días previos al partido. [16]

2. **Usar las apuestas previas de los usuarios con sus respectivos resultados de los partidos para las siguientes apuestas**. Si un usuario falló en sus apuestas, va a tener menos peso a la hora de decidir el resultado final. Encambio sí tuvo más aciertos, va a tener más peso

[16] Por ejemplo, si el tweet es el siguiente: *"vamos brasil, brasil, brasil, brasil"* cuenta solo un voto para Brasil

3. DESARROLLO

3.1. Generación del dataset de perfiles de Twitter etiquetados con el equipo de que es hinchas

Para poder estudiar cómo se comportan los hinchas de los equipos, necesitamos un ground truth de perfiles de twitter etiquetados con el equipo correspondiente por el cual el usuario es hincha. Nuestro primer objetivo fue armar el dataset. No sabíamos qué cantidad era la necesaria así que necesitábamos que nuestro procedimiento para encontrar hinchas pueda reconocer la mayor cantidad de hinchas posibles. Ademas cualquier método propuesto para resolver esto tiene un límite de hinchas ya que la cantidad de hinchas de los equipos es finita y depende del equipo. No es lo mismo buscar hinchas para un club como Boca Juniors que es reconocido a nivel mundial que para Arsenal de Sarandí.

3.1.1. Método utilizado

Intenamos varias formas hasta encontrar un método que nos sirva. En section 10.1 explicamos los intentos fallidos de generar el dataset.

3.1.1.1. ¿Qué miran los colaboradores en los perfiles para poder decidir el equipo por el cual el usuario es hincha?

Para lograr una mayor velocidad en etiquetar a los perfiles, nos pusimos a pensar en algún método computacional y pensamos que es lo que los colaboradores miran en los perfiles para poder decidir el equipo por el cual el perfil es hincha.

- Su biografía, screen_name, y name ya que en estos puede aparecer apodos de los cuadros de futbol
- Los tweets ya que es la forma la cual los usuarios se expresan, compartiendo publicaciones de otras cuentas, conversando con otros usuarios, o simplemente escribiendo tweets mencionando a jugadores o apodos del equipo
- Las cuentas que sigue el perfil ya que si sigue a más cuentas relacionadas de un equipo que el resto, tiene más chances de ser hincha de ese equipo
- Imagen de perfil y portada ya que pueden incluir imágenes relacionadas al club como por ejemplo imágenes del estadio, escudo del club, jugadores, directores técnicos

3.1.1.2. Nuestra solución para decidir el equipo por el cual el usuario es hincha

El recurso que encontramos para poder conseguir los hinchas fue utilizar los seguidores que sigan a una cuenta semilla pero los que cumplan todas las condiciones que mencionaremos. Notar que este método no tiene ni interacción humana ni ninguna validación humana de los perfiles como se estaba haciendo en los otras propuestas fallidas. Los pasos a seguir son:

1. Identificar cuentas semillas para cada equipo quitando jugadores. [1]

2. Buscamos los seguidores de esta cuenta y nos quedamos con los seguidores obtenidos del paso anterior que cumplan las siguientes condiciones necesarias [2]para etiquetarlo como un hincha reconocido de un equipo:

   a) Se encuentra un apodo [3] del equipo deseado en su biografía, screen_name o name de Twitter

   b) El seguidor realizó más RT a la cuenta oficial o cuentas del equipo deseado que a cualquiera de los otros equipos restantes posibles

   c) El seguidor conversó con la cuenta oficial o cuentas del equipo deseado mediante @ más que con cualquiera de los otros equipos restantes posibles

   d) El seguidor sigue a más cuentas del equipo que a cualquiera de los otros equipos restantes

[1] Esto se debe a que a los jugadores de fútbol los siguen hinchas de diferentes equipos. Por ejemplo, entre las posibles cuentas de River se encuentra @dale10oficial que es el perfil de Twitter del ex-jugador de River, Andrés D'Alessandro, que actualmente juega en el equipo Inter de Brasil. Además, Andrés también jugó en San Lorenzo. Por ende, hinchas del equipo Brasilero y de San Lorenzo lo pueden seguir, y nosotros queremos evitar estas posibilidades.

[2] Tener en cuenta que no siempre el equipo devuelta por la primera condición es correcto. Por eso exigimos que se cumplan también las siguientes 3 condiciones para eliminar estos casos. Por ejemplo, podemos tener las siguientes posibles biografías:

- *'No soy ninguna santa'*
- *'Vivo en Santa Fe'*
- *'Huracán te amo. Antes de ser cuerva, yo me muero'*
- *'Fan de romeo santos'*

Son etiquetados como hinchas de San Lorenzo ya que Santo y Cuervo son apodos de San Lorenzo y la herramienta no tiene en cuenta a los hinchas de Huracán. Sin embargo, en ninguna de las cuatro biografías se encuentra un posible hincha de San Lorenzo.

[3] Una aclaración importante es que los apodos de los diferentes equipos fueron seleccionados a mano, es decir, no se utilizó ningún algoritmo o herramienta para seleccionar los apodos sino que se ingresaron uno a uno por conocimiento previo de los apodos de los diferentes equipos encontrados en Internet. Algunos equipos comparten apodos o abreviaciones, como por ejemplo matador para Tigre y San Lorenzo y CAT para Talleres y Tigre. Decidimos eliminar estos apodos o abreviaciones para evitar confusión ya que capaz aunque no tengamos en cuenta a Tigre como un equipo a buscar hinchas, puede etiquetar a un hincha de Tigre como un hincha de San Lorenzo simplemente por el apodo. Mismo ejemplo con Talleres y Tigre con respecto a la abreviación.

Lo llamaremos de ahora en más requisitos del perfil de Twitter para ser etiquetado. Notar que tratamos de incluir lo mismo en nuestro set de reglas de experto que un colaborador humano puede verificar a mano salvo las imágenes ya que es algorítmicamente más complejo el análisis de tales.

Optamos este método porque necesitamos tanto mucha más cantidad de hinchas de los que conseguimos mediante los métodos anteriores como una cantidad uniforme de hinchas por equipo y además esta posibilidad tiene muy buena velocidad en conseguir los perfiles etiquetados. La desventaja que tiene la propuesta es que no existe confirmación por parte de un colaborador humano que la cuenta de Twitter se la etiquetó correctamente por ende pueden existir perfiles mal etiquetados pero la cantidad de estos que se puede dar es muy baja ya que sería muy paradójico tener apodos de un equipo como nombre de tu usuario, y tanto seguir como interactuar con más perfiles que también sean de ese mismo club pero hinchar por cualquier otro.

3.1.1.3. ¿Cómo verificamos que nuestro algoritmo está etiquetando correctamente las cuentas de Twitter?

Seleccionamos 200 perfiles de los hinchas reconocidos por el método descrito anteriormente y los repartimos en 12 colaboradores donde cada colaborador se le asignó un equipo. La tarea de tal era verificar que las 200 cuentas pertenecían al club que se le asignó. De esta forma estamos evaluando la precisión de nuestro algoritmo al conseguir fanáticos de los equipos, y obtuvimos 0.99 %, que nos indica que aparentemente está etiquetando correctamente a los hinchas. Durante esta experiencia, los colaboradores detectaron que en los hinchas de los equipos pueden aparecer perfiles de ex-jugadores o famosos de los cuales capaz se sabe el equipo por cual hincha. Por ejemplo, apareció la cuenta @angelcorrea32 etiquetado como hincha de San Lorenzo pero este ex jugador de San Lorenzo es confeso hincha de Rosario Central.

3.1.1.4. ¿Cuántos perfiles etiquetamos por equipo con nuestra solución?

Sin embargo para algunos clubes no logramos aumentar a una cantidad tal que se pueda ver el análisis de sentimiento así que los descartamos quedándonos con Boca, River, San Lorenzo, Racing, Independiente, Rosario Central y Newell's ya que la cantidad de hinchas reconocidos de estos equipos es superior a los 2.000

La cantidad de hinchas reconocidos por cada club se puede ver en la siguiente tabla

Club	Cantidad hinchas reconocidos
River	3597
San Lorenzo	3035
Racing	2783
Boca	2703
Newell's	2670
Independiente	2253
Rosario Central	2207
Talleres de cordoba	1398
Estudiantes de La Plata	646
Belgrano de Cordoba	577
Gimnasia de La Plata	479
Lanus	443

3.2. ¿Qué es una cuenta semilla?

Llamaremos cuenta semilla a un usuario de Twitter que es seguido por muchos hinchas de un mismo equipo y tiene como propiedad que trata de distinguir a los seguidores de su cuenta como hinchas del mismo club que la semilla con un valor alto en la métrica de precisión. Es decir, además de la cuenta oficial del club, existen un conjunto de cuentas que si un usuario las sigue significa que lo más probable es que el usuario sea hincha de ese equipo. Por ende, las cuentas semillas permiten discriminar a la audiencia.

Notar que la cuenta más seguida entre los hinchas de ese equipo no siempre los caracteriza como hinchas. Por ejemplo, pueden aparecer cuentas como @CFKArgentina, y @mauriciomacri que no caracterizan a hinchas de ningún equipo pero también puede aparecer una cuenta como @cuervotinelli que caracteriza a los hinchas de San Lorenzo pero Marcelo Tinelli también es seguido por hinchas de otros equipos ya que es un famoso, perdiendo el rasgo de caracterizar a los hinchas de San Lorenzo. A las cuentas semillas se les tiene que hacer la siguiente pregunta: ¿Qué tan probable es que los seguidores del perfil de Twitter sean hinchas de un único equipo en particular? Y la respuesta esperada es muy probable. De esta forma se aumenta la métrica de precisión al obtener perfiles de Twitter etiquetados con el club que simpatizan. Por ejemplo, con @cuervotinelli al realizarse esa pregunta se obtiene como respuesta poco probable pero al realizar la misma pregunta con @RiverEPMA se obtiene muy probable.

3.2.1. ¿Cómo se generan las cuentas semilla?

Lo ideal para obtener las cuentas semilla de un equipo en particular es utilizar todos los hinchas del equipo lo que resulta imposible. Por ende, seleccionamos 400 fanáticos de cada equipo con verificación humana de que cada perfil de la red social corresponda al equipo etiquetado. A estos los llamaremos perfiles de hinchas reconocidos a mano. En total reconocimos 8000 perfiles de Twitter de esta forma que pertenecen a 20 equipos.

Ranking	screen name	cantidad	Ranking	screen name	cantidad
1	MundoAzulgrana	392	26	afa	112
2	SanLorenzo	375	27	deboedovengocom	112
3	cuervotinelli	314	28	LosCuervosDePoe	112
4	MatiasLammens	293	29	San_Lorenzo	110
5	ElPipiRomagnoli	256	30	AngelCorrea32	109
6	sanlorenzotvnet	210	31	TiendaSoyCuervo	108
7	negroortigoza	202	32	aguerosergiokun	107
8	jabuffarini	188	33	MusicuervoSL	102
9	PICHIMERCIER	174	34	DebateSL	101
10	emmamasok	174	35	Mascherano	98
11	SanLorePrimero	170	36	FotosSL	97
12	schcasla	163	37	elcosodelapizza	97
13	mcaute7	158	38	Siempre_Ciclon	93
14	lacicloneta	157	39	SLOnline	92
15	TitoVillalba15	152	40	enzoka23	90
16	LocuraCuervaa	145	41	dflatorre	89
17	Argentina	143	42	PochoLavezzi	87
18	comunidadcuerva	143	43	PielCuerva	84
19	CASLABasquet	137	44	NotidSanLorenzo	83
20	matias_caruzzo	126	45	cuervamarce	82
21	BoedoAzulgrana	122	46	JPVarsky	82
22	TyCSports	122	47	caslajuveniles	82
23	DiarioCuervo	117	48	todonoticias	82
24	EzeCerutti	115	49	lea_alves	81
25	SebaTorricoOk	115	50	LeoNavarro42	81

Tab. 3.1: Top 50 de cuentas más seguidas por los 400 hinchas de San Lorenzo

Los 20 clubes de primera división del fútbol Argentino seleccionados fueron: Atletico Tucuman, Banfield, Belgrano de Cordoba, Boca Juniors, Colón de Santa Fe, Estudiantes de La Plata, Gimnasia de La Plata ,Huracán , Independiente, Lanús, Newell's Old Boys, Racing, River, Rosario Central, San Lorenzo de Almagro, Talleres de Córdoba, Temperley, Tigre, Union de Santa Fe, y Vélez Sarsfield.

Luego de obtener los perfiles de hinchas reconocidos a mano, nos fijamos por equipo cuáles son las 100 cuentas más seguidas por hinchas de ese equipo.

Ilustramos como quedan las 50 cuentas más seguidas por los 400 perfiles que reconocimos a mano como fanáticos de San Lorenzo en Table 3.1 mostramos el screen name del perfil y la cantidad de cuervos que siguen a esa cuenta. Vemos en Tabla 3.1 que entre los top 50 de cuentas seguidas por fanáticos de San Lorenzo, nos encontramos con perfiles como @cuervotinelli, @Argentina, @TyCSports, @afa, @aguerosergiokun, @Mas-

Club	San Lorenzo		Independiente		Newell's	
Ranking	screen name	cantidad	screen name	cantidad	screen name	cantidad
1	MundoAzulgrana	392	OrgulloRojoWeb	375	CANOBoficial	390
2	SanLorenzo	375	Independiente	354	MR11ok	267
3	cuervotinelli	314	InfiernoRojo	313	nachoscocco32	258
4	MatiasLammens	293	aguerosergiokun	252	infoleprosa	236
5	ElPipiRomagnoli	256	locoxelrojoweb	212	PatonGuzman	152
6	sanlorenzotvnet	210	DeLaCunaAlInf	210	cuervotinelli	132
7	negroortigoza	202	cuervotinelli	150	Gatoformica10	131

Tab. 3.2: TOP 7 en los equipos San Lorenzo, Independiente, y Newell's

cherano, @elcosodelapizza, @dflatorre, @PochoLavezzi, @JPVarsky, @todonoticias que no son propios del equipo. Si nos preguntamos qué tan probable es que los seguidores de cualquiera de estas cuentas sean hinchas de San Lorenzo la respuesta es poco probable ya que son cuentas que no están íntimamente relacionadas con el club. Por ende, no queremos tener a este tipo de cuentas como semillas del club San Lorenzo.

Para eliminar este tipo de perfiles de Twitter lo que hacemos es utilizar distintas métricas que le asignan valores a cada perfil, y nos quedamos con los que obtuvieron mayor poder discriminante. Además, utilizaremos la técnica de One vs All. Es decir, para obtener las cuentas semilla de un equipo compararemos sus 400 perfiles de hinchas reconocidos a mano contra el resto de perfiles de hinchas reconocidos a mano como simpatizantes de otro club.

En lo que resta de esta sección, llamaremos atributos a los 100 perfiles de Twitter que son seleccionados como las cuentas más seguidas por los 400 perfiles de hinchas reconocidos a mano en base a quienes siguen estos últimos. Por ejemplo, los 50 perfiles mostrados en Tabla 3.1 son atributos del cuadro San Lorenzo. Por ejemplo, el atributo @cuervotinelli podemos ver en Tabla 3.2 que figura en varios equipos en los primeros lugares del ranking.

Para calcular el valor de las métricas en los atributos, utilizamos los 400 perfiles de hinchas reconocidos a mano de los 20 equipo obteniendo un total de 8000 perfiles de hinchas reconocidos a mano. Armamos una tabla que llamaremos X, donde las columnas representan a los atributos mientras que las filas representan a los perfiles de hinchas reconocidos a mano. Completamos el valor de una celda en la fila F columna C, de la siguiente forma:

- Si el perfil F sigue al atributo C, con un 1
- Caso contrario, con un 0

Tambien armamos otras 20 tablas, que llamaremos Y[equipo], con una única columna pero con la misma cantidad de filas que la tabla anterior donde cada fila F será completada de la siguiente forma:

- Si el perfil F es del equipo de la tabla, completamos con el nombre del equipo

Usuarios \ Atributos	@MundoAzulgrana	@SanLorenzo	@cuervotinelli	@MatiasLammens	@ElPipiRomagnoli	@sanlorenzotvnet	@negroortigoza	@OrgulloRojoWeb	@Independiente	@InfiernoRojo	@aguerosergiokun	@locoxelrojoweb	@DeLaCunaAlInf	@CANOBoficial	@MR11ok	@nachoscocco32	@infoleprosa	@PatonGuzman	@Gatoformica10
@brianRomeo	✓	✓	✓	✓			✓				✓				✓				
@maroDiablo			✓					✓	✓	✓	✓		✓						
@juanLeproso			✓								✓			✓	✓	✓	✓	✓	✓

Tab. 3.3: Tabla X

Etiqueta
San Lorenzo
Otro
Otro

Tab. 3.4: Tabla Y de San Lorenzo

- Caso contrario, con "Otro"

Se puede ver en las tablas 3.3, 3.4, 3.5, 3.6 un ejemplo.

Ahora que tenemos creadas las tablas X e Y, podemos usar las métricas para ver qué atributos son los más característicos para cada equipo.

Usamos las siguientes métricas: InfoGainAttributeEval, CorrelationAttributeEval y GainRatioAttributeEval de weka mientras que forest ,mutual_info_classif, f_classif,chi2 de sklearn. Nos quedamos con los atributos seleccionados por tales, que fueron alrededor de 50 atributos por equipo.

Etiqueta
Otro
Independiente
Otro

Tab. 3.5: Tabla Y de Independiente

Etiqueta
Otro
Otro
Newell's

Tab. 3.6: Tabla Y de Newells

4. EXPERIMENTACION

4.1. Descripción del dominio de datos

Vamos a analizar el comportamiento de los hinchas de algunos clubes durante el Campeonato de Primera División 2016-17. Los clubes que utilizamos son Boca, River, Racing, Independiente, San Lorenzo, Newell's y Rosario Central. Descartamos el resto de los clubes ya que no logramos obtener suficientes hinchas.

En Table 4.1 se puede ver cuantos hinchas obtuvimos de cada equipo, la cantidad de tweets que obtuvimos de cada club y la cantidad de tweets promedio por hincha en cada uno.

Recordar que obtuvimos 26.601.818 tweets en nuestra base pero en esta base hay tweets de equipos que no utilizamos como por ejemplo de Talleres o Belgrano. Utilizando solo estos equipos, tenemos un total de 22.101.459 tweets.

En Figure 4.1 se puede ver la cantidad de Tweets por fecha en los días sábado y domingo.

La explicación de porqué hay más Tweets en las últimas fechas que en el comienzo, ya que lo intuitivo sería que sea una cantidad similar de Tweets por fecha ya que son siempre los mismos hinchas, es que la API de Twitter te permite devolver los últimos 3.200 Tweets de una cuenta. Como los Tweets se obtuvieron desde Marzo en adelante y hay perfiles que escriben muchos estados, se pierden parte de estos. Por ejemplo, si el perfil tiene 6.000 Tweets tan solo se puede acceder a los 3.200 Tweets más recientes siendo imposible conseguir los primeros 2.800 estados de esa cuenta. Por ende, los usuarios que escriben demasiados Tweets por día, no logramos obtener todos sus Tweets.

Notar que durante los días de partido se escribieron 3.367.195 tweets del total de 22.101.459, que representa un 15 %.

Utilizando nuestra herramienta de análisis de sentimiento obtuvimos 4.729.658 de tweets con su score_tag y polarity_term_list correspondiente que representan un 0.17 % de la cantidad total de Tweets en la base. El criterio que utilizamos para seleccionar a qué tweets realizar análisis de sentimiento fue si la fecha de creación de tal se encuentra

Equipo	Cantidad de hinchas	Cantidas de Tweets	Cantidad de tweets promedio por hincha
River	3597	4.521.143	1.256
San Lorenzo	3035	3.177.049	1.046
Racing	2783	3.947.921	1.418
Boca	2703	3.180.339	1.176
Newell's	2670	2.445.138	915
Independiente	2253	2.690.774	1.194
Rosario Central	2207	2.139.095	969

Tab. 4.1: Cantidad de Tweets por equipo

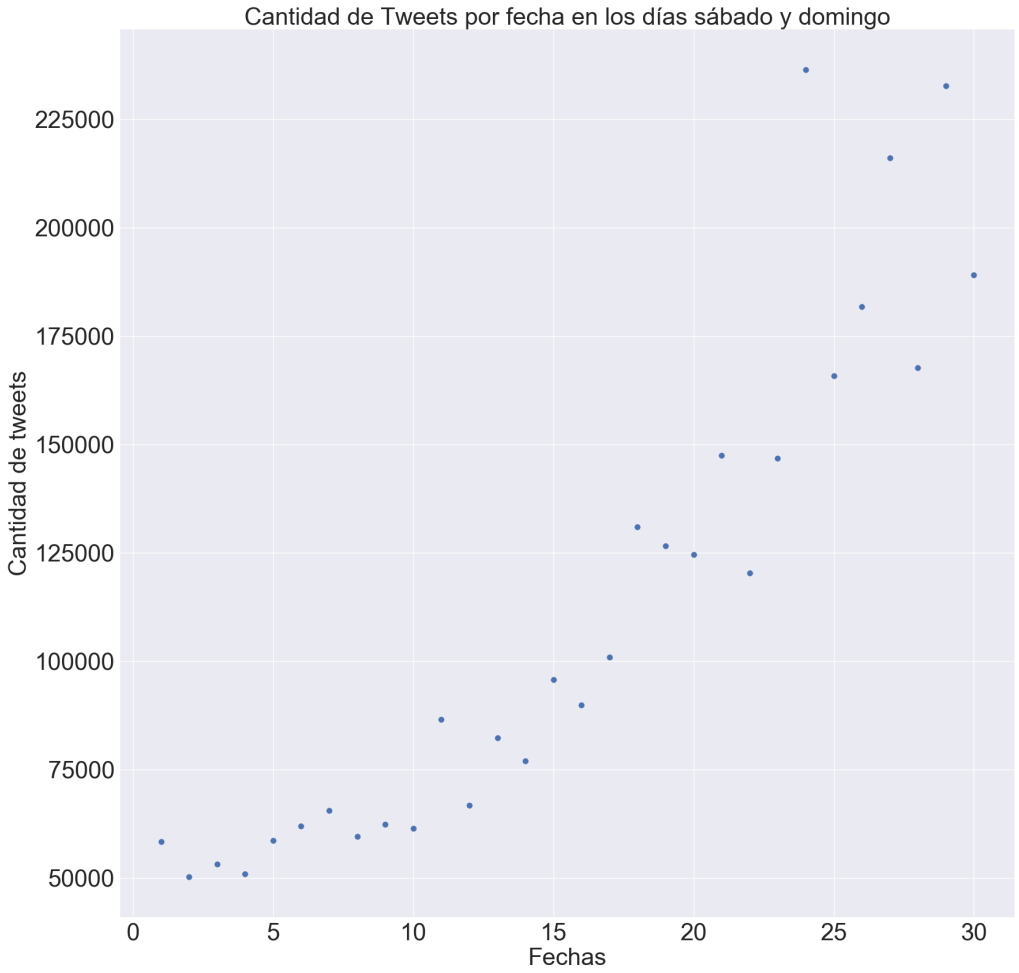

Fig. 4.1: Cantidad de Tweets por fecha en los días sábado y domingo

Equipo	Cantidad con sentimiento	Cantidad total	% sentimiento	% total
Boca	791.717	3.180.339	0.24 %	0.04 %
River	1.066.704	4.521.143	0.23 %	0.04 %
Independiente	544.147	2.690.774	0.20 %	0.038 %
Newell's	468.420	2.445.138	0.19 %	0.03 %
Racing	788.424	3.947.921	0.19 %	0.03 %
San Lorenzo	573.845	3.177.049	0.18 %	0.03 %
Rosario Central	402.902	2.139.095	0.18 %	0.028 %

Tab. 4.2: Cantidad de Tweets con sentimiento por equipo.

durante el día del partido, desde las 0 horas hasta 50 horas después de la finalización del partido. Decidimos esa franja para analizar cuánto tiempo le dura el envion animico al hincha producto del resultado del partido como por ejemplo la euforia o felicidad al ganar mientras que la tristeza o desazón al perder. Otro motivo por cual elegimos estos valores es para poder analizar la previa de los partidos no solo con la cantidad de Tweets sino también el sentimiento que expresan los hinchas en la red social.
En Figure 4.2 se puede ver cuántos tweets obtuvimos con sentimiento por fecha y en Table 4.2 cuantos tweets con sentimiento por equipo.

Notar que a pesar de que obtuvimos cantidades de Tweets con sentimiento muy distintas por equipo, tienen un porcentaje con respecto a la cantidad total de su equipo parecido.
Posiblemente los equipos que tienen más Tweets analizados obtienen una mejor precisión en los análisis que continúan ya que tienen más información para analizar. No es lo mismo tener 1.066.704 Tweets analizados en 30 fechas, dando un promedio de 35.556 Tweets por fecha, que tener 402.902 Tweets en la misma cantidad de fechas, dando un promedio de 13.430 Tweets por fecha.
Por otro lado la cantidad de Tweets por fecha con sentimiento mantiene la misma forma vista anteriormente sin análisis de sentimiento. Es decir, hay más tweets en las últimas fechas que en las primeras.
Utilizamos los 30 partidos de los 7 equipos mencionados anteriormente, generando un total de 210 enfrentamientos. Notar que en realidad son menos partidos, ya que cuando juegan entre sí los estoy contando en ambos pero como me interesa saber el resultado para cada equipo del partido, los contamos doble.
En 4.3, se puede ver cómo se distribuye el resultado de los partidos.

Lo que podemos ver en 4.3 es que aproximadamente la mitad de los partidos terminan siendo ganados, mientras que la otra mitad se divide entre partidos empatados y perdidos.

Además tenemos 105 partidos en condición de local y otros 105 partidos de visitante como era de esperarse.

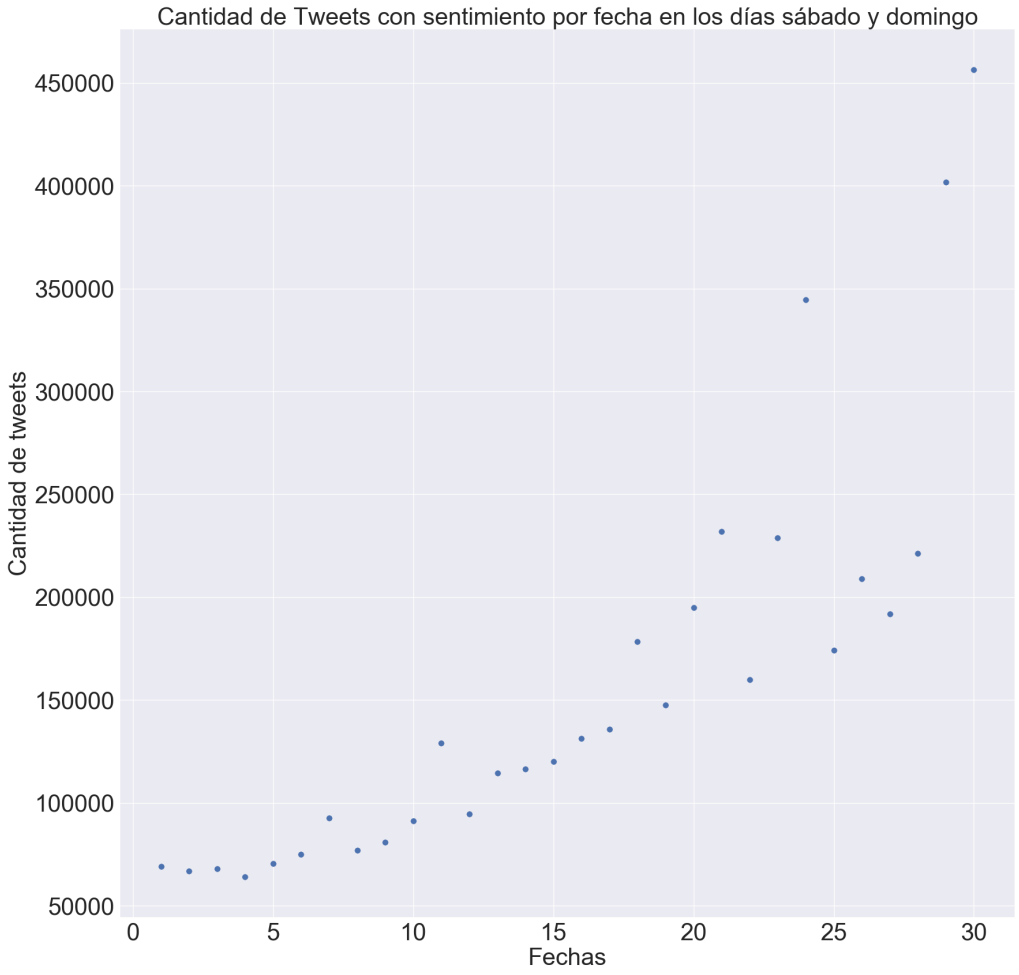

Fig. 4.2: Cantidad de Tweets con sentimiento por fecha en los días sábado y domingo.

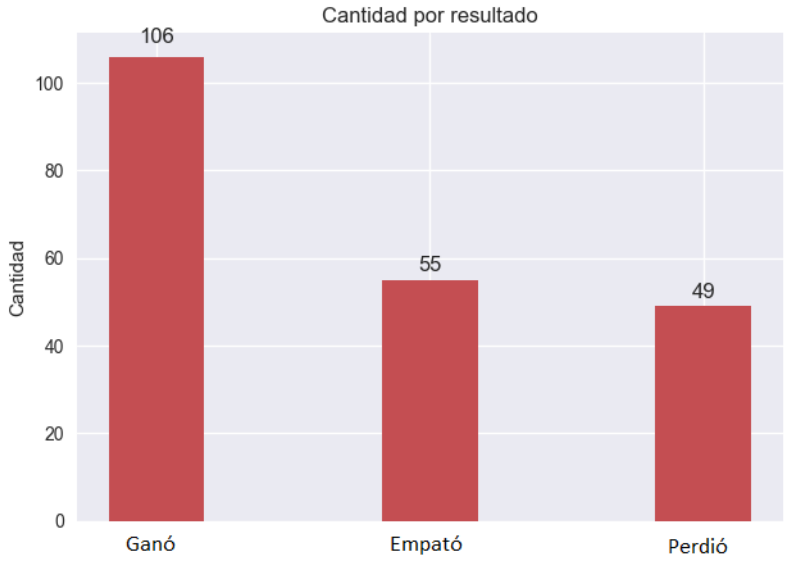

Fig. 4.3: Cantidad de partidos por resultado

4.2. Sobre los experimentos

4.2.1. Metricas

Para realizar los experimentos relacionados al humor de los hinchas utilizamos 3 métricas.

4.2.1.0.1. Metrica Facebook
Esta métrica se aplica a cada Tweet y se basa en la siguiente fórmula:

$$\frac{\#PalabrasPositivas}{\#TotalPalabras} - \frac{\#PalabrasNegativas}{\#TotalPalabras} \tag{4.1}$$

dónde #TotalPalabras es la cantidad de palabras que contiene ese tweet,
#PalabrasPositivas es la cantidad de palabras catalogadas positivas que figuran en ese tweet
mientras que #PalabrasNegativas es la cantidad de palabras catalogadas negativas que figura en el mismo tweet.

Las listas de palabras positivas y negativas son armadas de la siguiente forma:

- Se realizó el análisis de sentimiento a todos los tweets utilizando la herramienta de Meaning Cloud de análisis de sentimiento
- Se obtuvieron todas las palabras distintas que se encontraban en la Polarity_term_list

- Se cuenta todas las veces que aparece cada palabra como P+ o P y se las compara con la cantidad de veces que aparece esa misma palabra como N+ o N. En caso de que la primer cantidad sea mayor, se la asigna a la lista de palabras positivas. En el caso que la segunda cantidad sea mayor, se la asigna a la lista de palabras negativas.

Para calcular el humor del conjunto de hinchas con esta métrica en un momento

1. Calcular para cada Tweet el valor de la metrica Facebook

2. Calcular el promedio de los resultados anteriores

Notar que cuando este valor es positivo significa que hay más palabras positivas y cuando el valor es negativo significa que existen más palabras negativas.

La idea de la utilización de esta métrica surge al analizar [1]un artículo similar hecho por el grupo de data science de Facebook pero la adaptamos a los datos que tenemos disponibles.

4.2.1.0.2. Metrica Meaningcloud

A diferencia de la anterior, se aplica a un conjunto de Tweets que sucedieron en el mismo momento. Los pasos a seguir para obtener el valor de la métrica en un instante de tiempo son:

1. Utilizar para todos los tweets la herramienta de sentimiento de Meaning cloud

2. Cada tweet se le asigna un score_tag como resultado del paso anterior. Contar la cantidad de P o P+ que existen como cantidad de tweets positivos en ese instante. Contar la cantidad de N o N+ que existen como cantidad de tweets negativos en ese instante El valor de la métrica es la diferencia entre la cantidad de tweets catalogados como positivos con la cantidad de tweets etiquetados como negativos

Valor de la metrica =
#CantidadDeTweetsConEtiquetaPositiva - #CantidadDeTweetsConEtiquetaNegativa

Notar que cuando este valor es positivo significa que hay más tweets que el servicio de meaningcloud los etiqueto como P o P+ que N o N+ y cuando el valor es negativo significa que existen más tweets etiquetados con N o N+. En caso de que la diferencia sea 0, es que existe la misma cantidad, por ende el sentimiento en ese momento es neutral.

4.2.1.0.3. Metrica Meaningcloud Normalizada

Similar a la anterior pero se divide por la cantidad de Tweets utilizados generando un número entre -1 y 1.

[1] Link al artículo

Valor de la metrica =

$$\frac{\#CantidaddeTweetsconetiquetaPositiva - \#CantidaddeTweetsconetiquetanegativa}{\#CantidaddeTweetsconetiquetaPositiva + \#CantidaddeTweetsconetiquetanegativa} \tag{4.2}$$

4.2.2. Metodología para validar conclusiones

Para validar lo que a priori vemos en los gráficos, utilizamos el test T de Student [2] y el test de Mann-Whitney [3]para verificar que dos conjuntos de datos son significativamente distintos entre sí y además comparamos las medias de las muestras.

Para ambos test la hipótesis nula es que los dos conjuntos son iguales mientras que la hipótesis alternativa es que los dos conjuntos son distintos.

La diferencia entre ambos es que el primero pide que los conjuntos de datos analizados tengan una distribución normal mientras que el segundo no tiene este requisito.

El error de tipo 1 sucede cuando siendo la hipótesis nula cierta , se la rechaza. En este caso particular sería cuando los dos conjuntos de datos son iguales, rechazando que lo sean.

Llamamos nivel de significación del test a la probabilidad de error tipo 1. En particular, usamos un nivel de significación del test igual a 0.05.

Por otro lado, el p-value nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido si suponemos que la hipótesis nula es cierta.

Dicho esto, cuando el p-value es menor o igual al nivel de significación del test rechazamos la hipótesis nula mientras que cuando es mayor no podemos rechazarla.

Notar que se puede mover el nivel de significación del test para controlar el error de tipo 1.

[2] Student's t-test

[3] Mann–Whitney U test

5. RESULTADOS

A continuación se presentan los resultados de la experimentación realizada respecto al comportamiento del hincha argentino en la red social Twitter.

5.1. Análisis utilizando solo las cantidades de tweets

En esta sección no vamos a realizar análisis de sentimiento de los tweets, sino que vamos a analizar las cantidades de tweets por minuto.

5.1.1. ¿Los perfiles etiquetados son fanáticos del fútbol?

Nos consultamos si los perfiles etiquetados siguen el fútbol de primera división de Argentina. Nuestra hipótesis es que si ya que fueron etiquetados usando cuentas semillas relacionadas con ese tema.

Para verificar que estas cuentas estén relacionadas con el futbol de primera división de Argentina, comparamos la actividad minuto a minuto de estas durante los fin de semana donde hay partidos contra los que no tienen actividad. Como teníamos 30 fin de semanas con partidos de fútbol y 12 fin de semanas sin actividad, dividimos los resultados por estos valores para normalizar.

Lo que podemos ver en Figure 5.1 es que los días donde hay partidos de fútbol, tenemos más tweets por nuestros hinchas seleccionados así que podemos afirmar que la mayoría de las cuentas pertenecen a fanáticos que no solo se identifican por algún equipo sino que también lo siguen y comentan durante los partidos de fútbol ya que podría valer el caso de que siguen a un equipo pero en otro deporte. También lo que podemos notar en este gráfico es que la diferencia en cantidad de tweets ocurre en horarios donde se juegan los partidos de fútbol dando más validez a nuestro pensamiento.
Graficamos la cantidad de tweets desde 75 minutos antes que comience el partido a 75 minutos después de que termine el partido con una granularidad de 1 minuto y filtramos por los tres posibles resultados: ganar, empatar o perder. Es decir, nos fijamos para cada minuto cuántos Tweets hay que pertenecen a hinchas cuyo equipo terminó ganando su respectivo partido.

5.1.2. ¿Existe alguna relación entre la actividad de los hinchas en Twitter y el resultado del partido?

Nos consultamos si los hinchas tienen mayor actividad en la red social Twitter dependiendo el resultado del encuentro. Nuestra hipótesis es que si su equipo favorito gana su partido, twittean más a que si pierden o empatan.

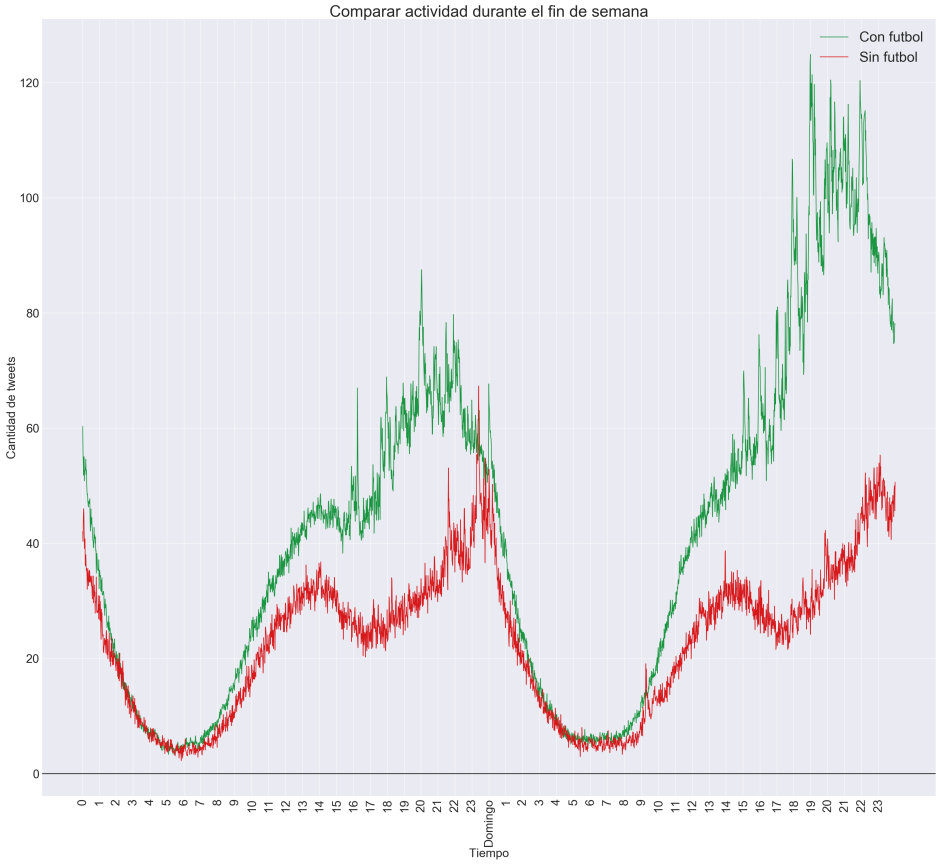

Fig. 5.1: Comparar fin de semanas con partidos de futbol contra los que no tienen partidos

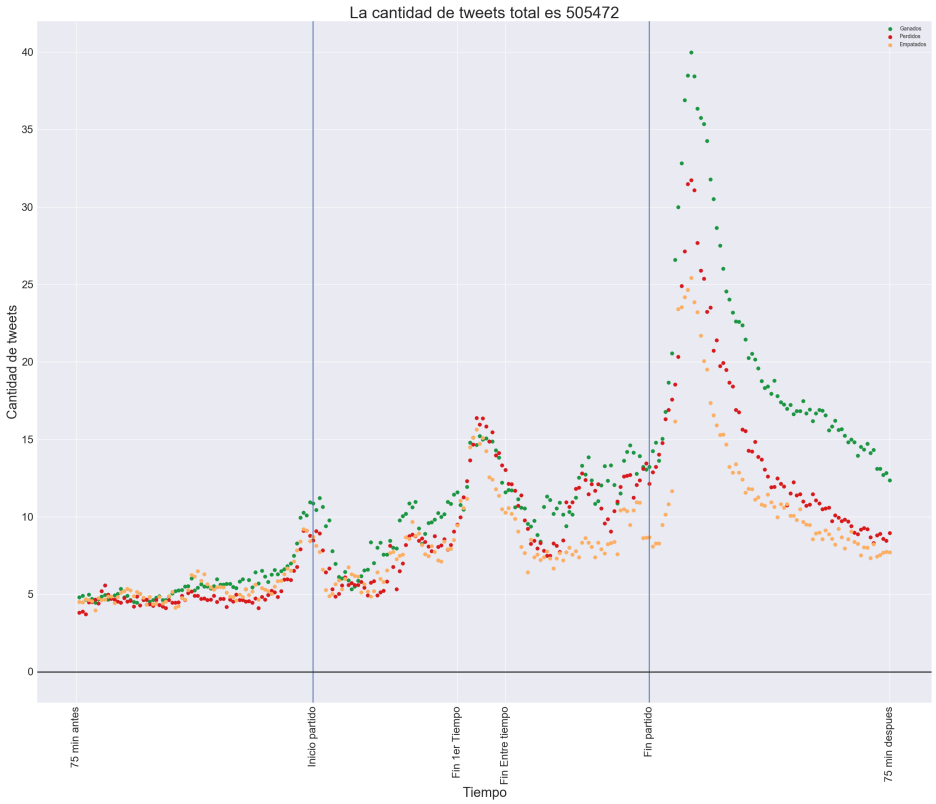

Fig. 5.2: Cantidad de tweets desde 75 minuto antes del inicio del partido hasta 75 minutos luego de la finalizacion filtrado por resultado

Valor de la media	En la previa	Durante el partido	Post partido
Ganado	934	1743	3344
Empatado	526	824	1174
Perdido	486	934	1441

Tab. 5.1: Valor de la media para la cantidad de tweets por periodo por resultado

Valor de la media	En la previa	Durante el partido	Post partido
Ganado vs Perdido	1.147e−23	9.952e−29	2.227e−22
Ganado vs Empatado	3.719e−23	1.164e−32	2.277e−25
Perdido vs Empatado	6.151e−5	0.001	7.644e−5

Tab. 5.2: Valor del p-value para la comparacion de valores por minuto de cantidades de tweets por periodo entre resultado

Lo que podemos observar en Figure 5.2 es que al ganar los simpatizantes tienen más actividad en la red social mientras que cuando pierden o empatan su actividad es similar y menor a la del triunfo. Esto último se nota muy claro al finalizar el encuentro, donde la cantidad de tweets de equipos que ganaron sus partidos no solo es mucho mayor sino que crece mucho más y esto creemos que se debe a que al finalizar el partido capaz los hinchas manifiestan sus sentimientos pero cuando su equipo gana muestran su festejos en Twitter incluyendo desde Tweets en agradecimiento al plantel, o felicidad por la victoria o bromas a hinchas del equipo rival. Lo que nos sorprendió del gráfico anterior es que también en la previa del partido hay más cantidad de tweets de los clubes victoriosos y pensamos que esto se debe a cómo está conformada nuestra base ya que tiene un sesgo en la cantidad de Tweets para equipos que terminaron ganando sus partidos.

También se puede ver que hay una subida en la actividad durante los entre tiempo pero no es un dato que nos interese profundizar más.
Una aclaración de todos los gráficos es que los momentos marcados de los partidos como el inicio y finalización tanto del primer tiempo como del segundo o goles o expulsiones son aproximaciones ya que los partidos pueden comenzar más tarde de lo previsto o tener más de 15 minutos de entretiempo. Por ejemplo en la figura anterior que señala el fin del partido debería estar más corrida a la derecha ya que se ve que la mayoría de los partidos no respetan los horarios establecidos.

5.1.3. ¿Existe alguna relación entre la actividad de los hinchas en Twitter y los momentos más importantes del partido?

Por lo visto en las últimas imágenes, se puede inferir cuándo un partido finalizó por la cantidad de Tweets. Dicho esto nos surgió la pregunta de si también se pueden inferir las situaciones mas importante de los partidos como los goles o expulsiones. A diferencia de la finalización del encuentro, estas situaciones pueden suceder en cualquier momento por ende para analizar esto graficamos para todos los equipos en las 30 fechas el minuto a minuto en cantidad de tweets.

Nuestra hipótesis es que al convertir un gol, se genera una gran cantidad de tweets festejando la situación mientras que al recibir un gol también una subida en cantidad de tweets lamentando tal situación pero mucho menor cantidad total.

En Figure 5.3 se puede ver el minuto a minuto en cantidad de Tweets para River en la fecha 18.

Logramos ver que no es tan claro como la finalización del partido pero se puede ver en la mayoría de los casos cuando un equipo marca un gol como los hinchas festejan en la red social mientras que hay un silencio cuando le convierten.

En estos gráficos una línea verde vertical significa que el equipo que nos interesa marcó un gol en ese momento mientras que la línea roja significa que le convirtieron un gol. Cuando la línea vertical es punteada y roja, significa que el equipo que nos interesa perdió un jugador debido a una expulsión mientras que si es verde el arbitro expulso a un jugador del rival.

Para eliminar la posibilidad de que por algún motivo los hinchas de distintos equipos escriban en el mismo minuto, nos fijamos qué porcentaje de todos los tweets de la base pertenecen al equipo en ese momento. Por ejemplo, podemos ver en el gráfico anterior de River que marca un gol al rededor del minuto 22 y que salta la cantidad de Tweets en el minuto 26. Por lo mencionado anteriormente, esto se debe a que no todos los partidos comienzan en la hora que están programados, así que podemos suponer por la cantidad de tweets que hay alrededor del minuto 26 que en realidad el gol fue un par de minutos después. Dicho esto, queremos ver que el aumento en la cantidad de Tweets se corresponda al gol en sí y no a un evento global que puede hacer Twittear a varios hinchas de distintos clubes como por ejemplo algún evento político o catástrofe natural.

Podemos ver en Figure 5.4 que se llega casi a un 50 % de los tweets escritos en ese momento por hinchas de River al momento de marcar el gol así que podemos asumir que es correcta nuestra hipótesis ya que el 50 % restante se divide en tweets de otros 11 equipos, que incluso uno de esos equipos puede ser el equipo rival.

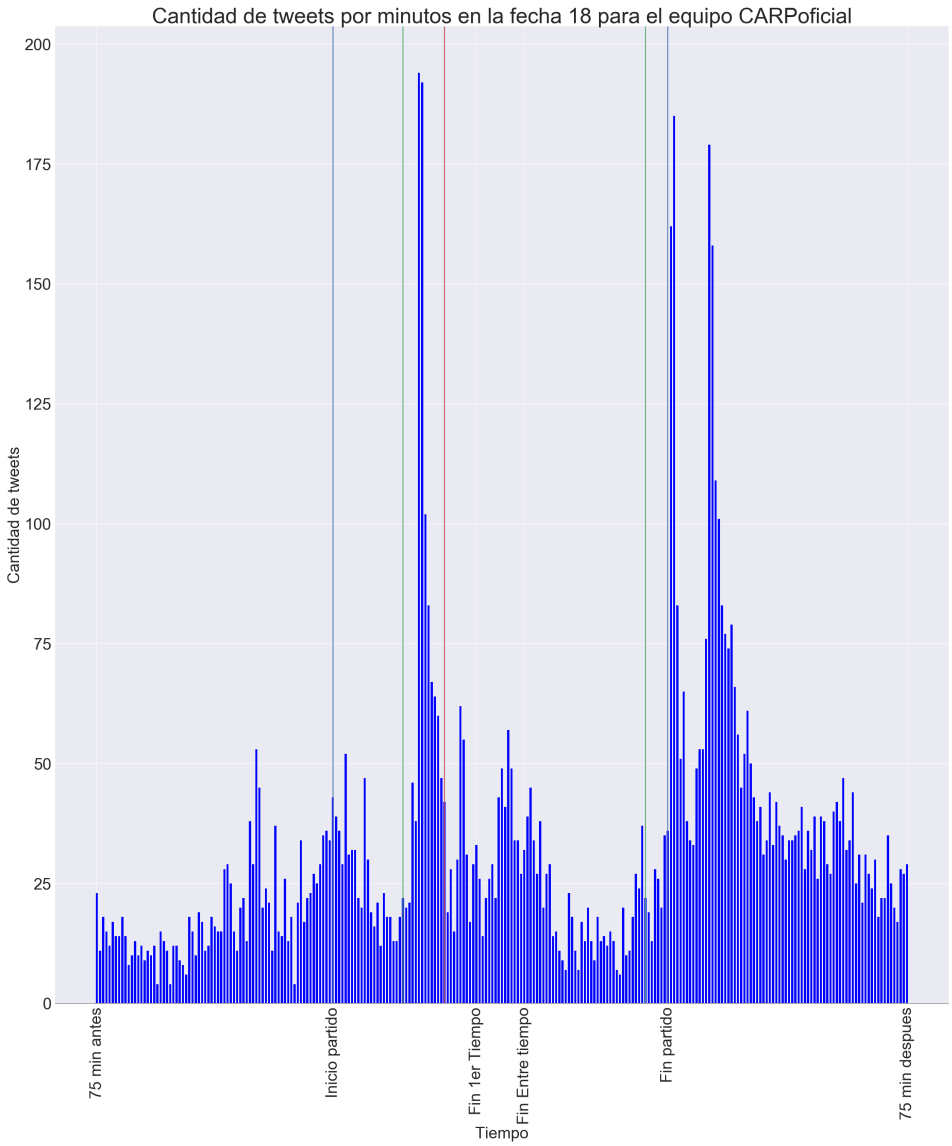

Fig. 5.3: Minuto a minuto en cantidad de Tweets para River en la fecha 18

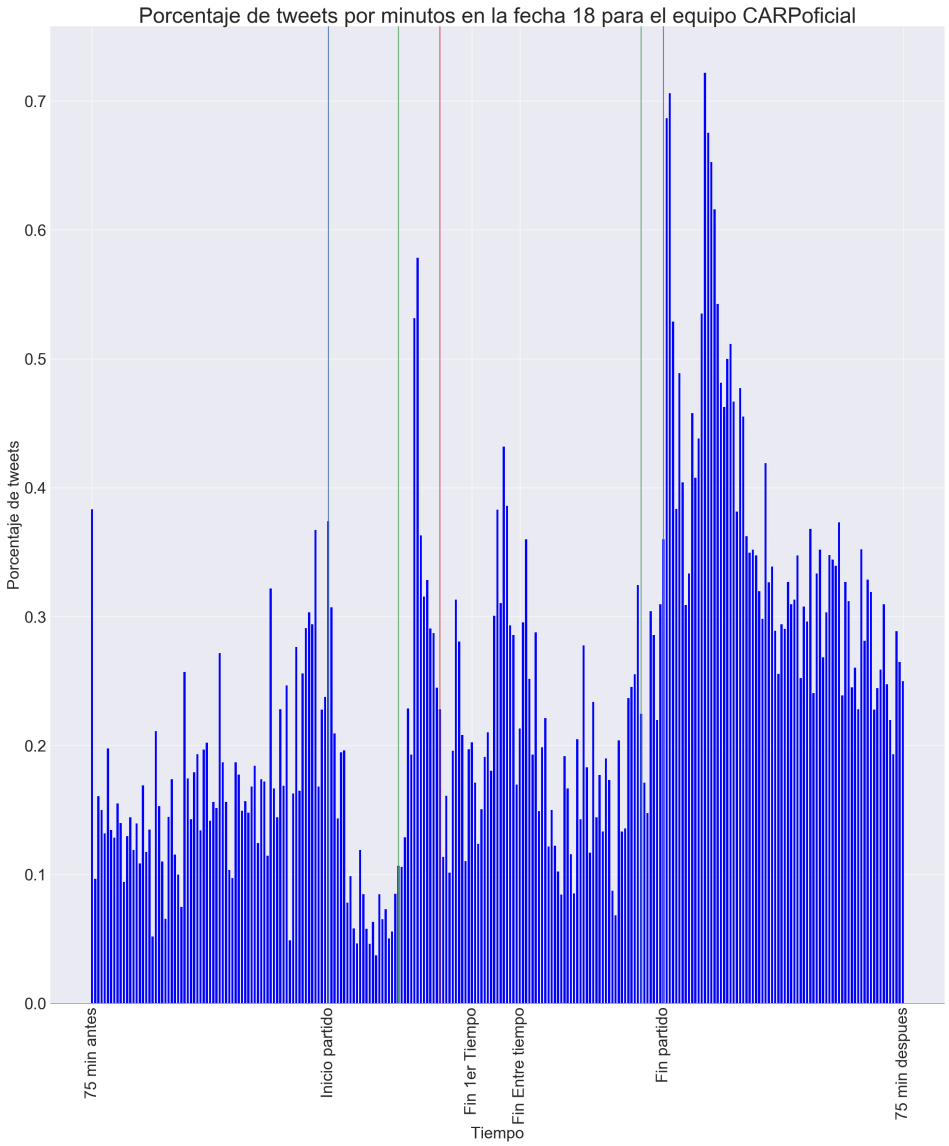

Fig. 5.4: Minuto a minuto en porcentaje de Tweets para River en la fecha 18

5.2. Análisis de sentimiento por partidos

El objetivo de esta sección es ir entendiendo cuál es el comportamiento de los hinchas durante y después de los partidos. **Las preguntas que nos intentamos responder es si reaccionan a los goles a favor y en contra y si el humor del hincha luego de la finalización del encuentro depende del resultado de tal**. En esta sección vamos a utilizar las métricas mencionadas anteriormente para analizar partidos pero de a un partido a la vez. Es decir, no vamos a analizar todos los partidos que cumplan una condición sino que vamos a examinar de a un partido. Como no realizamos el análisis de sentimiento para todos los equipos sino para un par, vamos a considerar partidos donde estos equipos se enfrenten entre sí así podemos ver la reacción minuto a minuto de ambos equipos en simultáneo para así poder observar primero cómo reaccionan los hinchas cuando su equipo marca un gol o le convierten.

5.2.1. El clásico de Avellaneda

En la fecha 11 se jugó en la cancha de Racing quien termina goleando por 3 a 0.

Lo que se puede apreciar en Figure 5.5 es que hasta el primer gol, los valores de las métricas de ambos eran similares pero luego del gol, los hinchas del equipo perdedor tuvieron la mayoría de los valores negativos mientras que los del ganador fueron positivos.

Para Figure 5.5 se utilizaron 7.201 tweets.

5.2.2. Boca vs Racing

En la fecha 12 Boca reciba a Racing que llegaba de ganar su clásico. El partido lo arrancó ganando los Xeneizes por 3 a 0 pero los de la academia descontaron para ponerse 3 a 2. Faltando 10 minutos para la finalización del encuentro, los locales convirtieron el último gol.

Lo que se puede ver en Figure 5.6 es que hasta el primer gol de Boca, ambos equipos tienen un sentimiento similar. Luego los hinchas de Racing pasan a tener un sentimiento con valores negativos producto del 3 a 0 parcial hasta que llega su primer gol que pasan a tener algunos valores positivos en ambas métricas cuando antes no tenían. Para este gráfico se utilizaron 7.149 tweets.

5.2.3. El superclásico

En la fecha 13 se jugó en la cancha de River donde el equipo visitante arrancaría ganandolo por la mínima diferencia a los 13 minutos. A pesar de que el local logró revertir el marcador antes de que finalice el primer tiempo, terminó perdiendo el partido ya que en el segundo tiempo Boca convirtió tres goles. Este partido está muy bueno para analizar ya que tiene varios giros en el marcador ya que en un momento lo arranca ganando

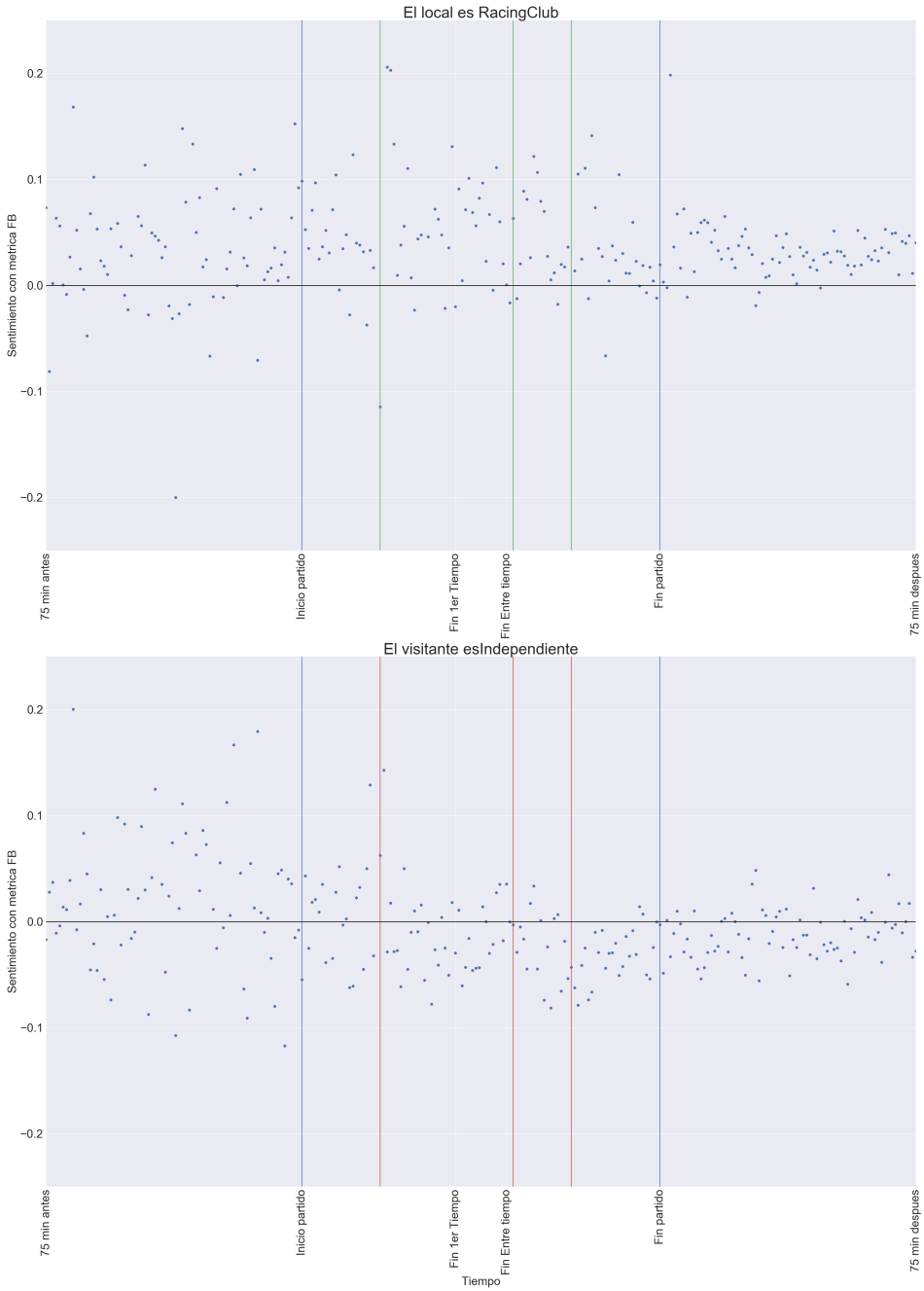

Fig. 5.5: Minuto a minuto del partido Racing vs Independiente de la fecha 11 utilizando la metrica Facebook

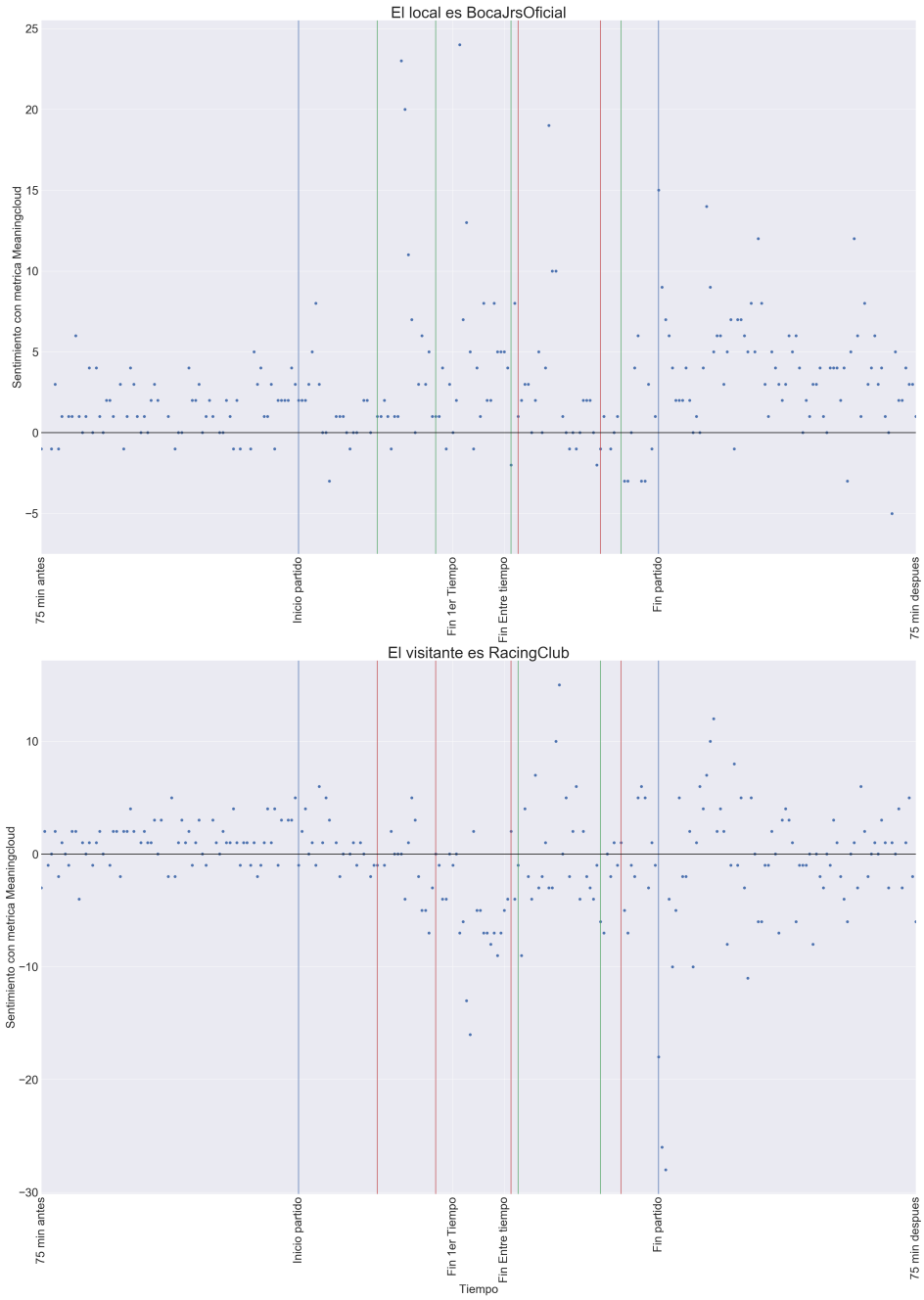

Fig. 5.6: Minuto a minuto del partido Boca vs Racing de la fecha 12 utilizando la metrica MeaningCloud

Boca, luego lo pasa a ganar temporalmente River para por último que lo gane el primero.

En Figure 5.7 se puede ver cómo los hinchas son influenciados por el resultado transitorio del partido. La primera situación es cuando Boca arrancó ganando, se puede ver como la mayoría de los valores de ambas métricas para River son negativas mientras que para Boca la mayoría son positivas. La segunda situación es cuando River da vuelta el marcador, también da vuelta lo que sucede en los gráficos de las métricas ya que la mayoría de los valores pasan a ser positivos para el conjunto de Núñez y negativos para el conjunto de la Boca. La última situación es cuando Boca termina dando vuelta el marcador siendo el resultado final, se vuelve a la primera situación. Para este gráfico se utilizaron 12.145 tweets.

5.2.4. Conclusiones de analizar por partido

Luego de analizar estos casos tenemos la intuición de que **el estado de ánimo en la red social se ve afectado por los goles que transcurren durante los encuentros especialmente cuando hay cambios en el resultado siendo estos ganar, empatar o perder**.
Otro resultado que podemos ver en los gráficos anteriores es que **al finalizar los encuentros, los equipos que ganan sus partidos tienen la mayoría de los valores de la métrica positivos mientras que al perder terminan con la mayoría de los valores negativos**.
Esto nos llevó a comparar el sentimiento medido por las tres métricas mencionadas anteriormente en todos los partidos filtrando por resultado.

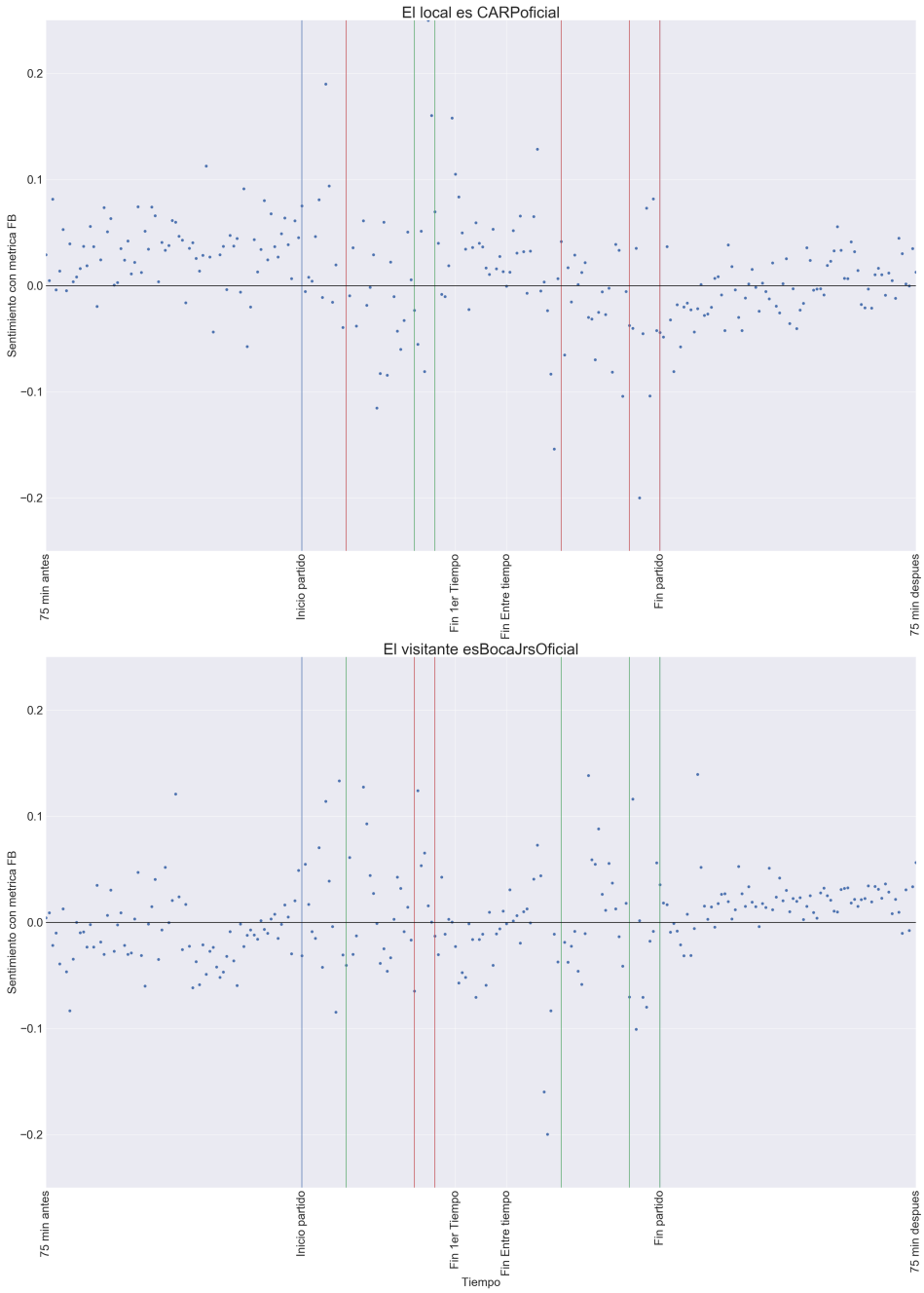

Fig. 5.7: Minuto a minuto del partido River vs Boca de la fecha 13 utilizando la metrica Facebook

Valor de la media	En la previa	Durante el partido	Post partido
Ganado	0.02	0.026	0.027
Empatado	0.014	0.010	0.005
Perdido	0.009	0.002	-0.001

Tab. 5.3: Valor de la media por periodo filtrado por resultado utilizando la metrica Facebook

Valor del p-value	En la previa	Durante el partido	Post partido
Ganado vs Perdido	4.041e−13	1.378e−67	3.150e−26
Ganado vs Empatado	2.929e−5	2.579e−27	2.557e−25
Perdido vs Empatado	5.418e−5	4.99e−13	1.715e−7

Tab. 5.4: Valor del p-value comparando por resultado filtrando por periodo utilizando la metrica Facebook

5.3. Análisis de sentimiento por resultado

El objetivo de este apartado es ver **cómo influyen los resultados de los partidos en los estados de ánimo de los simpatizantes. No solo de forma inmediata, sino cuánto tiempo les dura esta influencia**.

5.3.1. ¿Se puede ver una diferencia en el humor de los hinchas filtrando por resultado del partido?

Juntamos todos los partidos que finalizaron con una victoria y calculamos para cada minuto el valor en cada métrica. Luego realizamos lo mismo con los partidos cuyo resultado fueron un empate y por último los que el marcador decretaron una derrota.

Lo que obtuvimos en Figure 5.8 y Figure 5.9 es una separación muy clara desde que comienzan los partidos con respecto al humor de los hinchas. Los simpatizantes de los clubes que ganan sus partidos tienen valores más positivos a diferencia de los que pierden o empatan sus encuentros que tienden a ser similares aunque los que pierden

Valor de la media	En la previa	Durante el partido	Post partido
Ganado	145.013	272.801	312.078
Empatado	64.552	50.575	23.539
Perdido	31.539	10.226	-20.315

Tab. 5.5: Valor de la media por periodo filtrado por resultado utilizando la metrica MeaningCloud

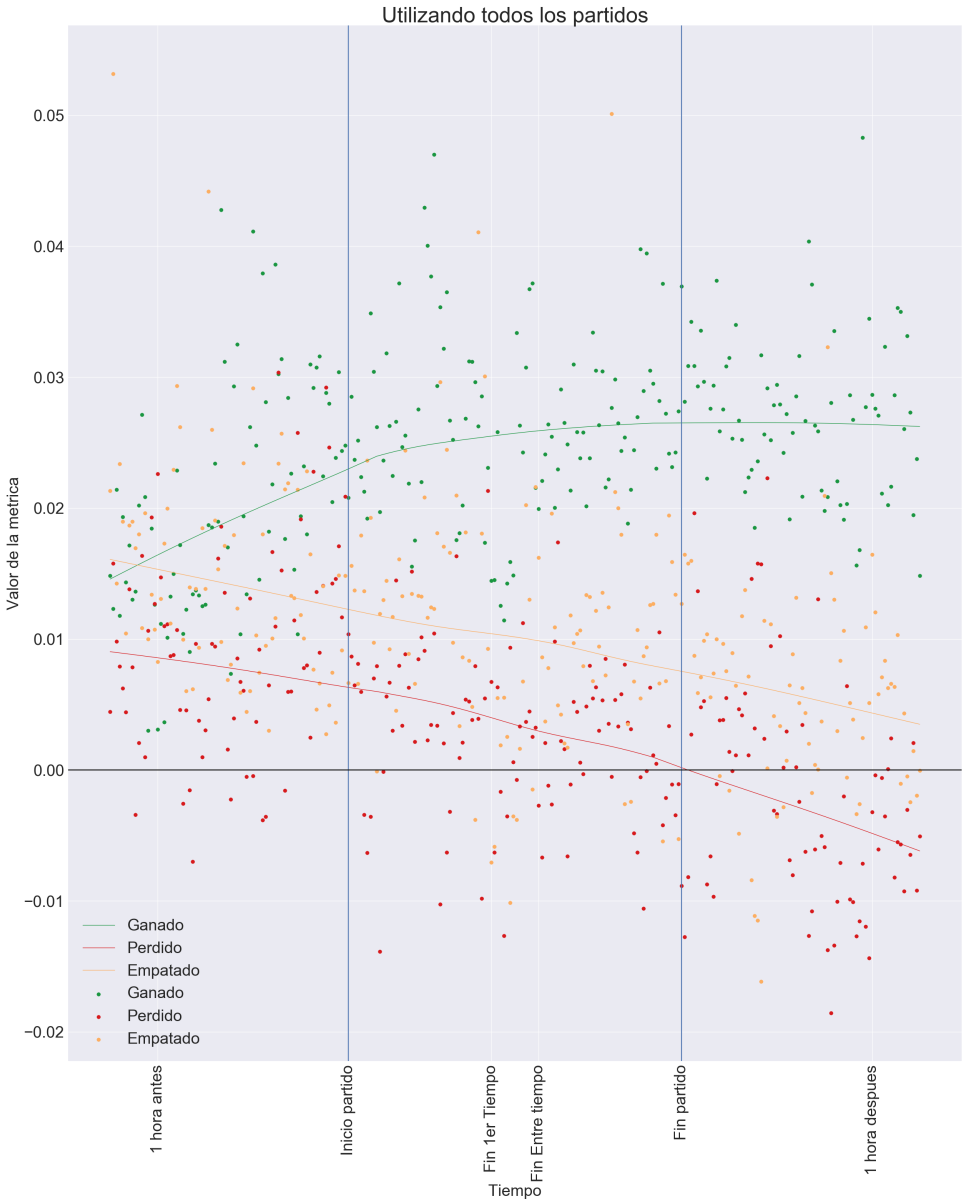

Fig. 5.8: Minuto a minuto desde 75 minutos antes del comienzo del partido hasta 75 minutos después de la finalizacion del mismo filtrado por resultado utilizando la métrica Facebook

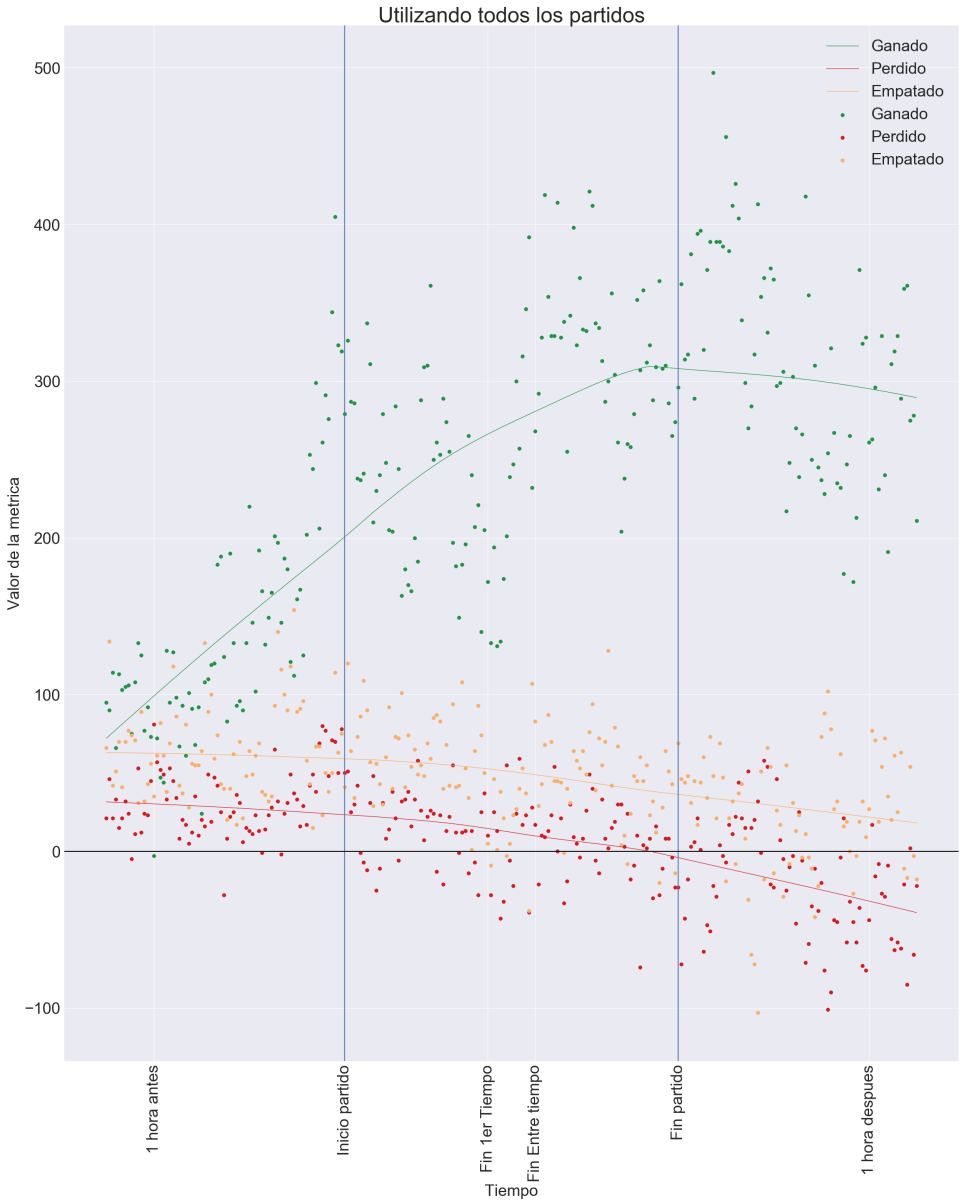

Fig. 5.9: Minuto a minuto desde 75 minutos antes del comienzo del partido hasta 75 minutos después de la finalizacion del mismo filtrado por resultado utilizando la métrica MeaningCloud

Valor del p-value	En la previa	Durante el partido	Post partido
Ganado vs Perdido	2.279e−23	6.333e−72	2.415e−66
Ganado vs Empatado	1.941e−15	2.784e−65	9.642e−27
Perdido vs Empatado	8.084e−12	9.091e−21	9.223e−11

Tab. 5.6: Valor del p-value comparando por resultado filtrando por periodo utilizando la metrica MeaningCloud

Valor de la media	En la previa	Durante el partido	Post partido
Ganado	85.671	128.103	178.868
Empatado	39.447	26.613	7.144
Perdido	13.513	3.254	2.078

Tab. 5.7: Valor de la media por periodo filtrado por resultado utilizando la metrica MeaningCloud mirando solo partidos de local

suelen tener un humor más negativo. En caso de mirar los 30 minutos previos al comienzo de los encuentros para las métricas de Meaning Cloud, se puede observar también la separación anteriormente mencionada.

En los gráficos anteriores usamos 1.009.685 tweets.

5.3.2. ¿Se puede ver una diferencia en el humor de los hinchas filtrando por resultado del partido en condición de local?

En los gráficos anteriores se encuentran todos los partidos que terminan con una victoria pero sin filtro por triunfo en condición de local o de visitante. Obtuvimos imágenes similares con las mismas conclusiones filtrando también por localia. Por ejemplo en el gráfico Figure 5.10 están todos los partidos jugados de local filtrados por resultado con la métrica Meaning Cloud utilizando 486.243 tweets

Dejaremos para section 9.1 un estudio más profundo de esta cuestión.

Valor del p-value	En la previa	Durante el partido	Post partido
Ganado vs Perdido	9.463e−26	1.017e−69	7.378e−51
Ganado vs Empatado	4.41e−17	1.760e−58	9.608e−27
Perdido vs Empatado	7.045e−15	9.820e−17	0.013

Tab. 5.8: Valor del p-value comparando por resultado filtrando por periodo utilizando la metrica MeaningCloud mirando solo partidos de local

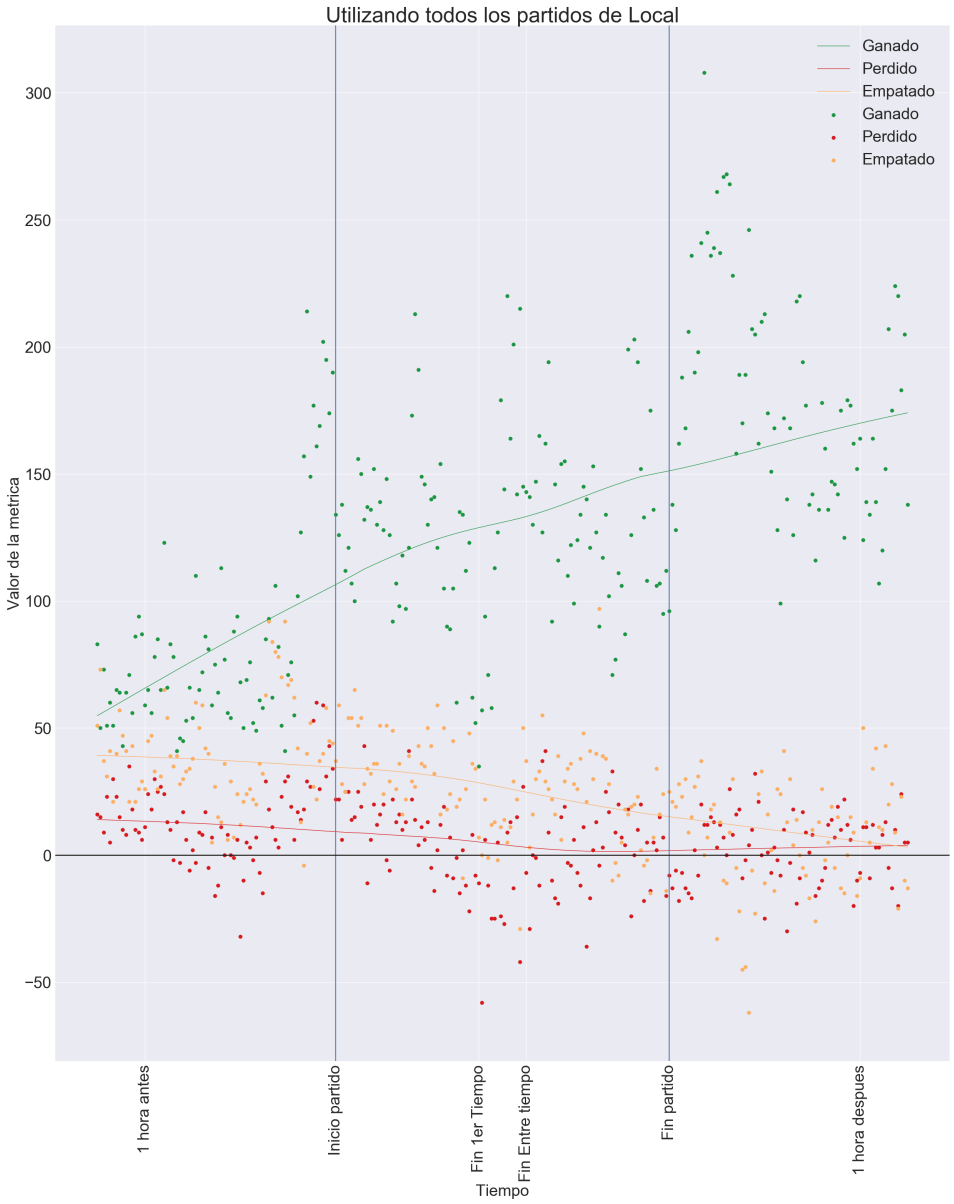

Fig. 5.10: Minuto a minuto desde 75 minutos antes del comienzo del partido hasta 75 minutos después de la finalizacion del mismo filtrado por resultado utilizando la métrica MeaningCloud en partidos en condición de local

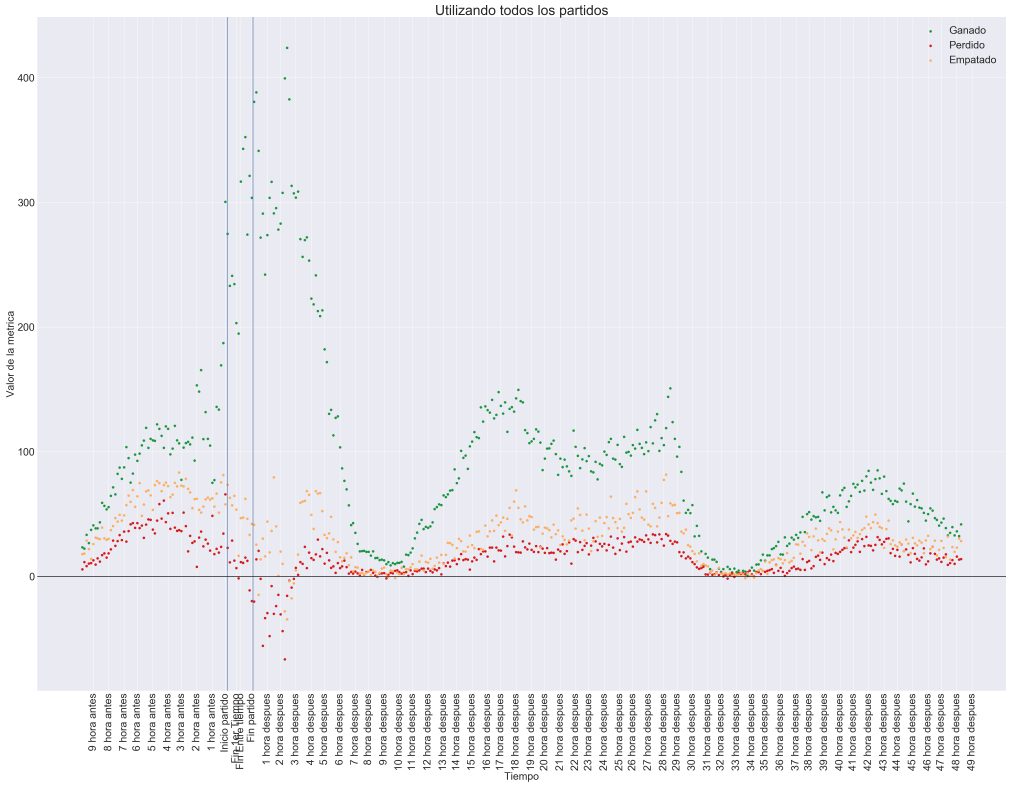

Fig. 5.11: Minuto a minuto desde 9 horas antes del comienzo del partido hasta 48 horas después de la finalizacion del mismo filtrado por resultado utilizando la métrica MeaningCloud

5.3.3. ¿cuánto tiempo les dura esta influencia?

Una pregunta que nos surgió relacionado a los posibles resultados de los partidos fue **¿cuánto tiempo les dura el impacto por el resultado a los hinchas**. Es decir si el equipo por cual soy simpatizante ganó, ¿cuánto tiempo me dura la felicidad luego de la victoria? Y si perdió, ¿cuánto tiempo me dura la tristeza luego de la derrota? Nuestra hipótesis es que duran tan solo 6 horas ya que al día siguiente el hincha se olvida del resultado. Otra pregunta que nos resultó interesante realizar es **¿cómo es el estado de ánimo previo al partido de los hinchas?**. Nuestra hipótesis es que 8 horas antes del partido todos los hinchas tienen un humor similar ya que aún no están influenciados por la excitación del partido.

En Figure 5.11 podemos responder las últimas dos preguntas. Este gráfico es similar a Figure 5.8 y Figure 5.9 pero la única diferencia es que comienzan desde 10 horas antes del comienzo del partido y finalizan 48 horas después de la finalización del encuentro utilizando 6.616.344 tweets.

Se puede ver en Figure 5.11 que nuestra primera hipótesis es incorrecta ya que **el**

impacto por el resultado dura aproximadamente 30 horas ya que luego la diferencia entre el humor de los distintos resultados es pequeña o no existe ya que se cruzan. Creemos que este resultado se debe a que dura el dia siguiente entero, que serían 24 horas, y las horas que le resten al día del encuentro para finalizar que calculamos que serán aproximadamente 6 horas ya que los equipos seleccionados suelen jugar desde las 18 horas en adelante.
Notar que apenas termina el partido, la diferencia de humor entre los hinchas de los equipos que ganaron sus cotejos contra el resto de los hinchas no solo es muy clara sino que es el momento que mayor diferencia existe, lo cual tiene mucho sentido ya que tienen el envión anímico del resultado del partido muy fresco. Con el paso de las horas esta diferencia se va achicando a tal punto de que la diferencia en algunos deja de existir.
Además se puede notar que hay 3 separaciones del humor post partido en forma de escalera.Es decir, el primer escalon recibe mas el impulso por el resultado del partido que los dos siguientes pero el segundo escalón también recibe más impulso que el tercer y último escalón. El primer escalón arranca desde apenas termina el partido hasta 6 horas después, luego el segundo desde 12 horas a partir de que termina el partido a 30 horas desde el mismo momento y el último escalón desde 38 horas de la finalización del encuentro durado aproximadamente 10 horas. Creemos que estos escalones se forman debido a que hay un freno en la actividad en la red social debido al horario ya que es tarde y los hinchas están durmiendo.
Por otro lado, se puede ver como **8 horas antes del comienzo del partido los hinchas tienen humores similares sin importar como luego finalicen sus respectivos encuentros. Nos pareció interesante que desde 3 horas antes del partido ya se puede notar una diferencia en el humor de los hinchas filtrados por el resultado del partido**. Esto lo vamos a tener en cuenta para luego predecir los resultados de los partidos.

Parecido a lo que nos sucedió con Figure 5.8 y Figure 5.9 obtuvimos imágenes similares con las mismas conclusiones filtrando también por localia.

5.4. Análisis de sentimiento por equipo

Realizamos un estudio similar al hecho en las secciones 5.3 Análisis de sentimiento por resultado y 9.1 Análisis de sentimiento por localia, pero visto por equipo.

Evidentemente, los resultados deberían ser los mismos ya que el anterior se basaba en todos los equipos y nosotros suponemos que la conducta de los hinchas de todos los equipos son similares. Es decir, un hincha de Boca se va a comportar igual que uno de River.

5.4.1. ¿Se puede apreciar las mismas conclusiones con todas las métricas mirando los resultados anteriores por equipo?

Lo que notamos en el análisis por resultado por equipo es que se llegan a las mismas conclusiones aunque los gráficos obtenidos no son tan claros ya que al tener tweets de un solo equipo en lugar de 7 equipos genera una menor cantidad de tweets logrando una precisión menor.

Además como no tenemos la misma cantidad de tweets en todos los equipos sino que tenemos equipos con cantidades muy distintas debido a cómo está conformada nuestra base de tweets, hay equipos que cuestan aún más poder apreciar las conclusiones en los gráficos.

Una conclusión que llegamos en Análisis de sentimiento por resultado es que el humor de los hinchas es influenciado por el resultado del partido durante 30 horas desde la finalización del encuentro y que 8 horas antes del comienzo del partido los hinchas tienen humores similares sin importar como luego finalicen sus respectivos encuentros.

En Figure 5.12 y Figure 5.13 podemos ver para el equipo Newell's como varía el humor de sus hinchas desde 10 horas antes del comienzo del partido hasta 48 horas después de la finalización del encuentro utilizando 510.998 tweets. La única diferencia entre estos gráficos es que Figure 5.12 corresponde a la métrica Meaning cloud mientras que Figure 5.13 corresponde a la métrica de Facebook.

En Figure 5.13 no se puede ver las conclusiones obtenidas utilizando todos los equipos, y esto se debe que cada punto graficado tiene menos Tweets entonces la métrica de Facebook no termina de tener la suficiente información como para llegar al valor regularizado. Sin embargo, en el gráfico Figure 5.12 se pueden ver los resultados deseados ya que utiliza **la métrica Meaning cloud que no es tan sensible a la cantidad de tweets**.

5.4.2. ¿Los fanáticos de River festejan más la victoria de local o de visitante?

En el 9.1 Análisis de sentimiento por localia llegamos a la conclusión de que se obtiene mejor valor de la métrica post partido cuando el equipo gana de local que de visitante.

Sin embargo, notamos que en el equipo River no sucede eso en Figure 5.14.

El pvalue de comparar los valor de la métrica meaning cloud en local contra visitante en partidos ganados de River es $8.540e{-}26$.

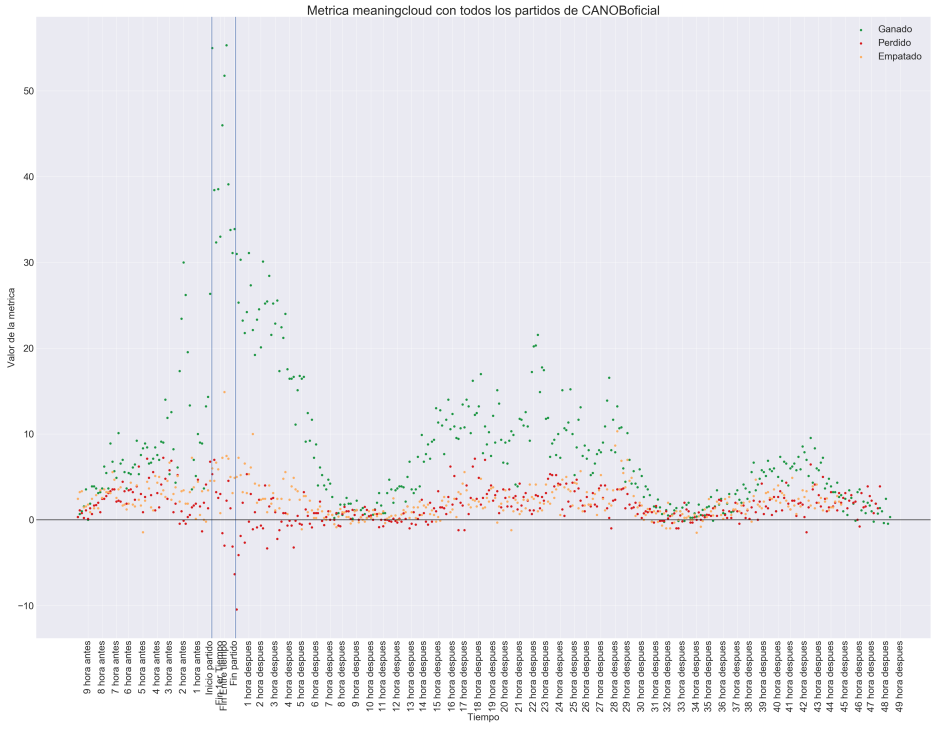

Fig. 5.12: Minuto a minuto desde 9 horas antes del comienzo del partido hasta 48 horas después de la finalizacion del mismo filtrado por resultado utilizando la métrica MeaningCloud en partidos del club Newell's

Valor de la media	Local	Visitante
Media	27.733	46.554

Tab. 5.9: Valor de la media para la metrica Meaningcloud filtrado por localia para partidos ganados de River

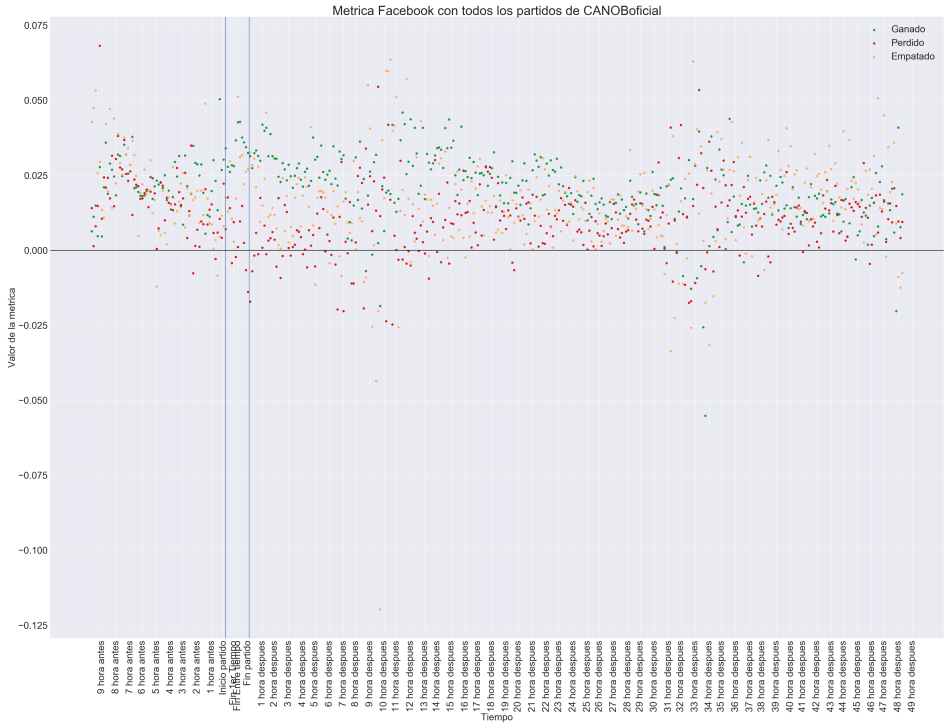

Fig. 5.13: Minuto a minuto desde 9 horas antes del comienzo del partido hasta 48 horas después de la finalizacion del mismo filtrado por resultado utilizando la métrica Facebook en partidos del club Newell's

Fig. 5.14: Minuto a minuto de los partidos ganados de River utilizando la metrica MeaningCloud filtrando por localia

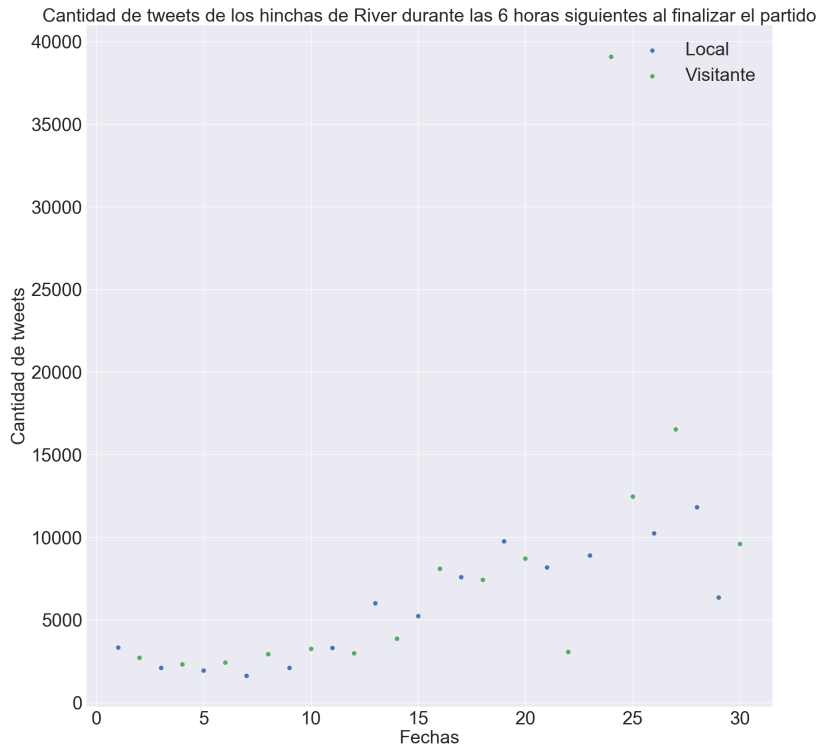

Fig. 5.15: Cantidad de tweets durante las 6 horas siguientes al finalizar el partido por fecha en el club River

Suponemos que tiene que ver con que ganó su clásico de visitante generando una gran cantidad de tweets en las horas siguientes. En Figure 5.15 se puede ver la cantidad de tweets durante las siguientes 6 horas luego de que finaliza el partido, y se puede ver que en la fecha 24, donde River le ganó 3 a 1 de visitante a Boca , no solo se obtiene la mayor cantidad de Tweets escritos sino por un amplio margen. En la fecha 24 se obtiene 39.085 Tweets mientras que la siguiente fecha con más tweets se consiguen apenas 16.536, que representa tan solo un 42 % de la primera cifra.

Valor de la media	Boca	River
Media	-0.036	0.054

Tab. 5.10: Valor de la media para la metrica Meaningcloud Normalizado filtrado por equipo para partidos con derrota de River y Boca

5.5. Análisis de sentimiento de a dos equipos

En esta sección comparamos de a dos equipos. Vamos a comparar

- Boca vs River
- Independiente vs Racing
- Newell's vs Rosario Central

Los seleccionamos de tal forma así podemos comparar los más grandes de cada ciudad. **El objetivo es poder encontrar si existe algún comportamiento a destacar entre las hinchadas de las rivalidades clásicas**. Por ejemplo, a pesar de una derrota los hinchas de River tienen mejor humor que los de Boca o los fanáticos de Independiente festejan más las victorias que su rival de Avellaneda.

En esta parte optamos por utilizar para la visualización de los resultados la métrica Meaning cloud normalizada ya que tenemos una buena cantidad de tweets como para asegurarnos que converge al valor final.

5.5.1. Boca vs River

Lo que notamos comparando los dos equipos más grandes de Argentina en Figure 5.16 es que **en las derrotas, los hinchas de Boca se ponen de peor humor que los de River**. Para ver esto utilizamos 3 derrotas con 38.461 tweets para los Xeneizes mientras que 6 caídas utilizando 55.066 tweets para los de núñez.

El pvalue de comparar los valor de la métrica meaning cloud normalizada de todos los partidos perdidos de Boca contra los partidos perdidos de River es $2.086e{-}16$.

5.5.2. Independiente vs Racing

Lo que notamos comparando los dos equipos más grandes de Avellaneda en Figure 5.17 es que **pese a perderlos hinchas de Racing, tienen mejor humor que los de Independiente al obtener el mismo resultado**. Para ver esto utilizamos 5 derrotas con 41.652 tweets para los del rojo mientras que 9 caídas utilizando 67.441 tweets para los de la academia.

Pensamos que lo ocurrido en las derrotas se debe a que la academia al perder 9 veces se acostumbró a la derrota, mientras que los hinchas del rojo al perder 5 veces a lo largo de un año no se acostumbraron a tal sensación y además los rivales de sus derrotas eran

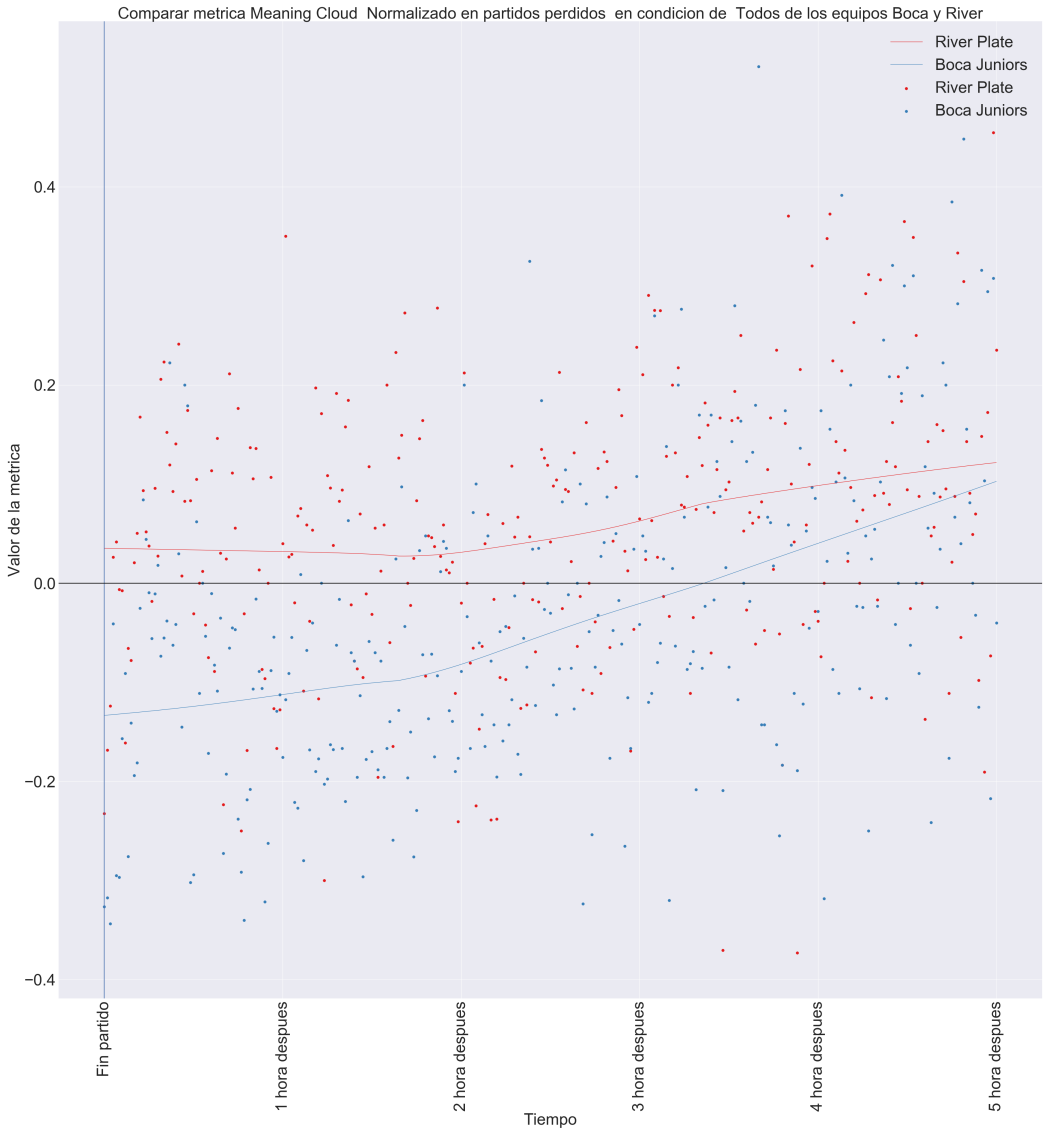

Fig. 5.16: Comparar post partido entre los equipos Boca y River en los partidos perdidos utilizando 93.527 tweets

Valor de la media	Racing	Independiente
Media	-0.011	-0.076

Tab. 5.11: Valor de la media para la metrica Meaningcloud Normalizado filtrado por equipo para partidos con derrota de Racing y Independiente

Valor de la media	Newell's	Rosario Central
Media	0.245	0.045

Tab. 5.12: Valor de la media para la metrica Meaningcloud Normalizado filtrado por equipo para partidos con derrota de Newell's y Rosario Central

más importantes ya que 3 de las 5 derrotas de Independiente fueron a manos de equipos como Racing, San Lorenzo, Independiente generando mayor bronca en los hinchas ya que no lograron ganar los clásicos mientras que Racing perdió 2 de sus 9 caídas contra equipos como Boca y Independiente y las 7 caidas restantes contra equipos chicos como por ejemplo Atlético Rafaela, Olimpo de Bahía Blanca, Unión de Santa fe, Temperley, entre otros.

El pvalue de comparar los valor de la métrica meaning cloud normalizada de todos los partidos perdidos de Racing contra los partidos perdidos de Independiente es 2.095e−8

5.5.3. Newell's vs Rosario Central

Para los equipos de Rosario notamos una actitud similar ante la derrota como la victoria. Sin embargo, notamos que **los hinchas de Rosario Central suelen tener peor humor luego de un empate** en Figure 5.18 Para ver esto utilizamos 7 partidos empatados con 11.161 tweets para los de Newell's mientras que 11 pardas utilizando 33.979 tweets para los de Rosario Central. Notar que a pesar de que los 7 partidos empatados de Newell's representen un 63 % de la cantidad total de partidos empatados por Rosario Central, tan solo tiene un 32.8 % de los tweets. Es decir, que ante los malos resultados los hinchas de Central suelen expresar más sus sentimientos en la red social.

El pvalue de comparar los valor de la métrica meaning cloud normalizada de todos los partidos perdidos de Racing contra los partidos perdidos de Independiente es 5.981e-30

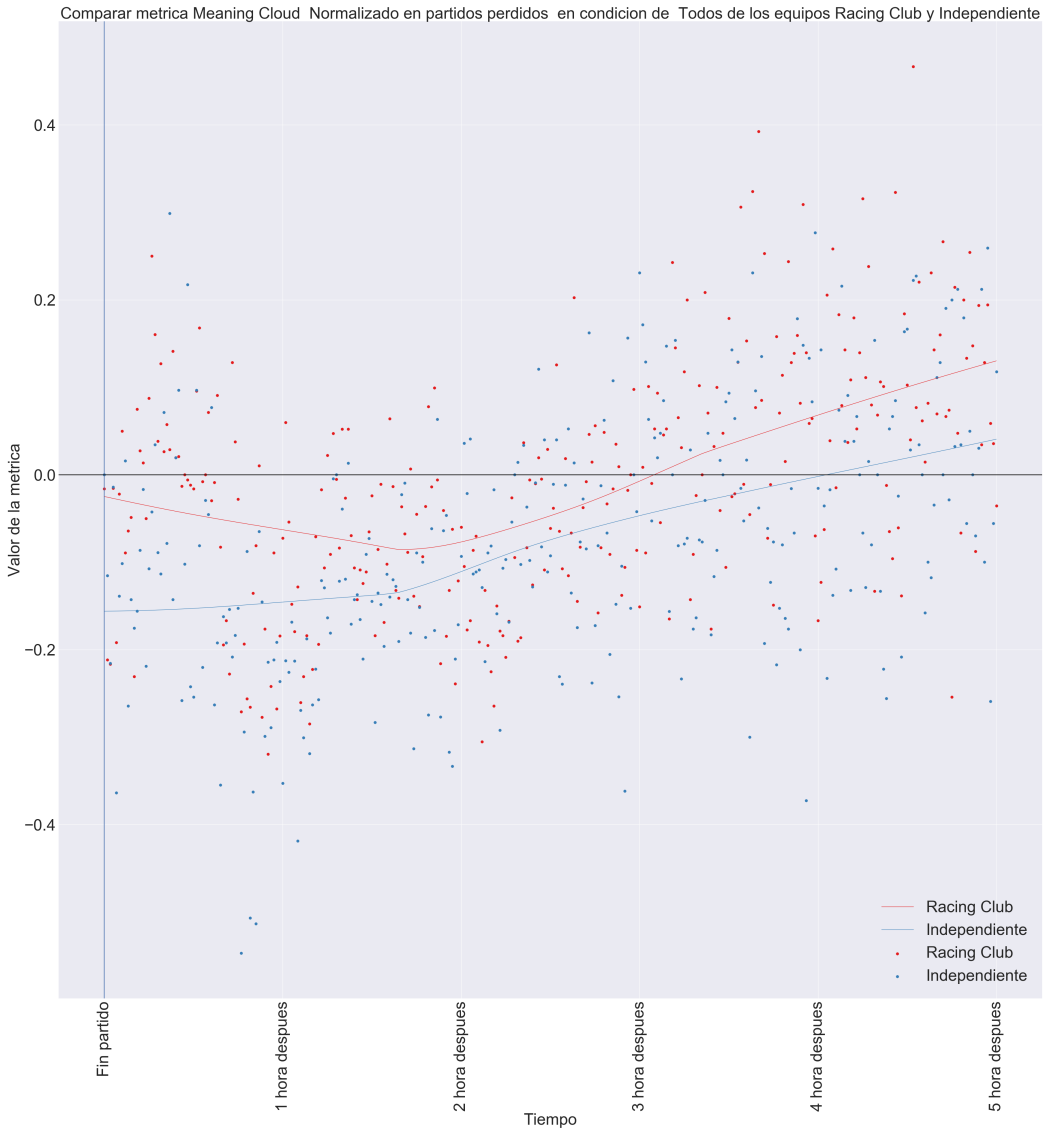

Fig. 5.17: Comparar post partido entre los equipos Independiente y Racing en los partidos perdidos utilizando 109.093 tweets

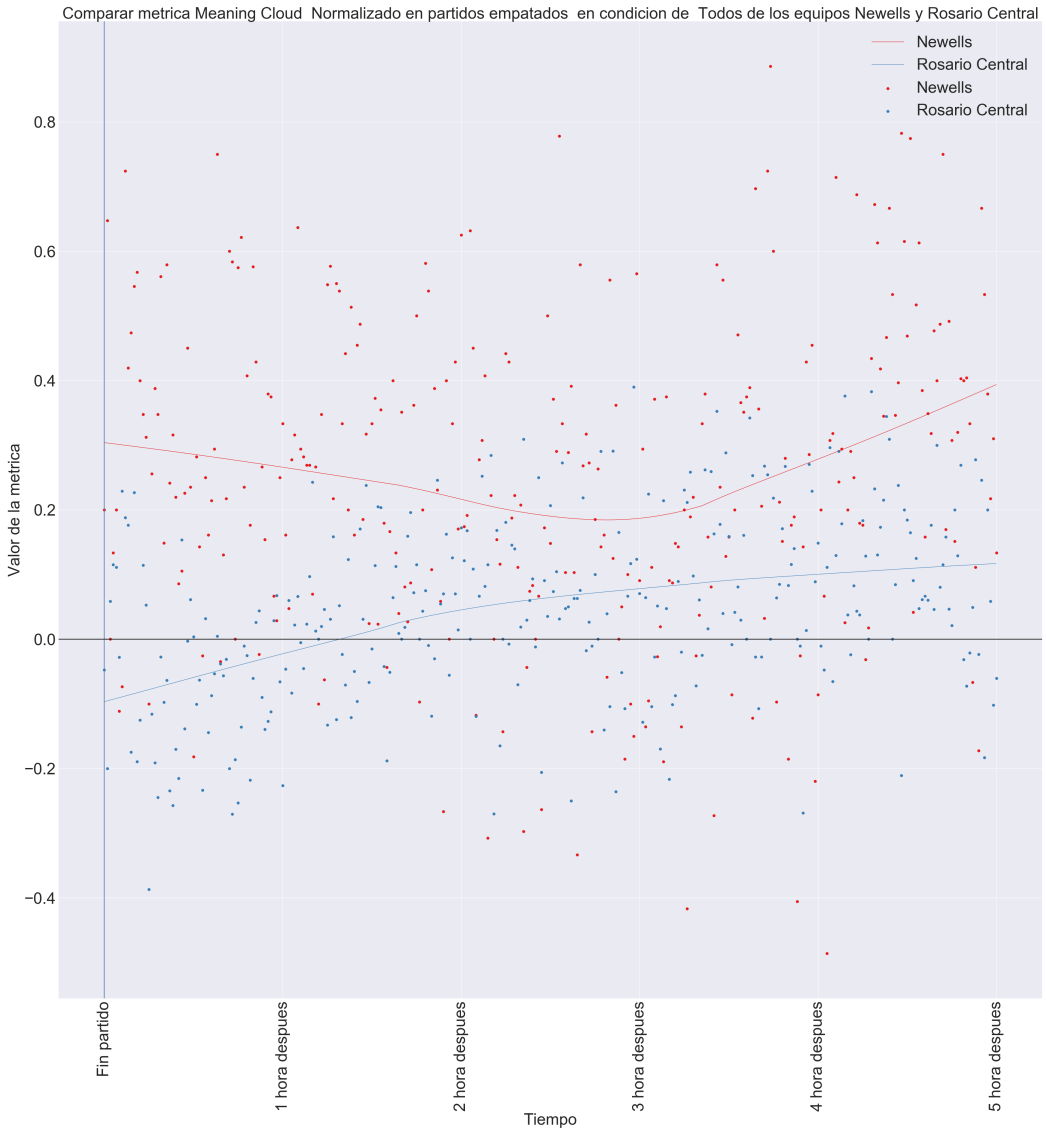

Fig. 5.18: Comparar post partido entre los equipos Rosario Central y Newells en los partidos empatados utilizando 45.140 tweets

Valor de la media	Todos salvo partidos clásicos	Solo clasicos
Media	0.225	0.263

Tab. 5.13: Valor de la media para la métrica Meaningcloud Normalizado en partidos ganados en condición de local filtrado por partidos no clásicos vs clásicos

5.6. Análisis de sentimiento de partidos clásicos contra el resto de los partidos

Existen partidos llamados clásicos que son cuando juegan equipos que tienen entre sí una rivalidad muy fuerte. Los clásicos analizados en esta sección son los siguientes:

- Boca contra River. En la fecha 13 Boca ganó en el Monumental mientras que River ganó en la fecha 24 en la Bombonera. Es decir, cada equipo ganó de visitante.
- Racing contra Independiente. En la fecha 11 Racing ganó en el Cilindro mientras que en la fecha 24 Independiente ganó en el Libertadores de América. Es decir, cada equipo ganó de local
- Newell's contra Rosario Central. En la fecha 7, Newell's ganó en el Gigante de Arroyito mientras que Rosario Central ganó en la cancha Marcelo Bielsa en la fecha 24.
- San lorenzo le ganó los dos clásicos a Huracán en las fechas 9 y 24, jugando de local y visitante respectivamente.

Notar que estos equipos fueron los mismos que utilizamos en la sección de Análisis de sentimiento de a dos equipos salvo San Lorenzo y Huracán.

5.6.1. ¿Los hinchas muestran más felicidad al ganar un clásico que el resto de los partidos?

La primera pregunta es si los hinchas muestran más felicidad al ganar un clásico que el resto de los partidos tanto de local como de visitante. Nuestra hipótesis es que los hinchas muestran más felicidad al ganar un clásico.

En Figure 5.19 se puede ver cómo se cumple nuestra hipótesis de local al finalizar los partidos. Obtuvimos los mismos resultados en condicion de visitante.

Para Figure 5.19 de ganados de local utilizamos 231.689 tweets, dónde 26.658 tweets corresponden a los clásicos.

El pvalue de comparar los valor de la métrica meaning cloud normalizada de todos los partidos ganados sin incluir los partidos clásicos contra los partidos clásicos ganados en condición de local es 9.388e−6.

Fig. 5.19: Comparar post partido en partidos ganados en condicion de local filtrado por clásico o no

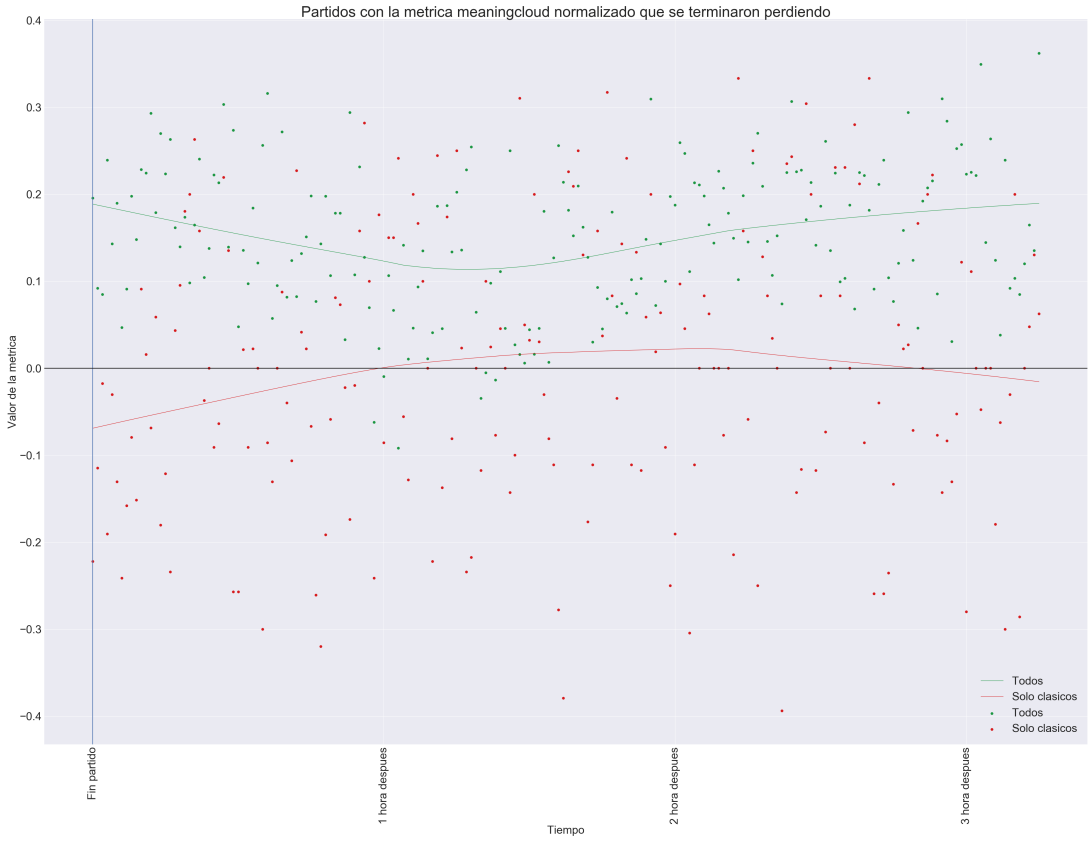

Fig. 5.20: Comparar post partido en partidos perdidos en condición de local filtrado por clásico o no

5.6.2. ¿Los hinchas muestran más angustia al perder un clásico que el resto de los partidos?

La segunda pregunta es si los fanáticos demuestran más angustia al perder un clásico que cualquier otro partido en ambas localias. Nuestra hipótesis es que los fanáticos se angustian más al perder un clásico.

En Figure 5.20 se puede ver cómo se cumple nuestra hipótesis de local al finalizar los partidos. Obtuvimos los mismos resultados en condicion de visitante.

Para Figure 5.20 de perdidos de local utilizamos 105.824 tweets, dónde 45.558 tweets corresponden a los clásicos.

El pvalue de comparar los valor de la métrica meaning cloud normalizada de todos los partidos perdidos sin incluir los partidos clásicos contra los partidos clásicos perdidos en condición de local es 6.782e−23.

También nos hicimos las mismas preguntas pero por equipo y obtuvimos los mismos

Valor de la media	Todos salvo partidos clásicos	Solo clásicos
Media	0.149	-0.001

Tab. 5.14: Valor de la media para la métrica Meaningcloud Normalizado en partidos perdidos en condición de local filtrado por partidos no clásicos vs clásicos

resultados. Por ejemplo, los seguidores de Boca festejan mas ganarle a River que a cualquier otro equipo pero una derrota contra el club de Núñez los angustia más que cualquier otra derrota.

6. MODELO DE PREDICCIÓN

6.1. ¿Qué buscamos predecir?

Nos propusimos armar modelos que puedan predecir 3 posibles etiquetas, las cuales son:

1. Llamaremos a este tipo de etiqueta resultado del partido y **consiste en si el equipo gana o empata o pierde su partido**. Para esto, utilizamos las siguientes etiquetas

   a) 0 cuando gana

   b) 1 cuando empata

   c) 2 cuando pierde

2. **La diferencia de gol**. En caso de ser positiva, significa que el equipo gana su encuentro, mientras que en caso de ser negativa significa que el equipo pierde su partido por esa diferencia. Por último, una diferencia de gol igual a 0 significa que termina en empate. Notar que es similar al anterior pero con más granularidad

3. **Cantidad de goles marcados en el partido**. No importa quien los marco, nos interesa la cantidad total de goles convertidos. Es decir, la suma de los goles convertidos tanto del local como del visitante.

Vamos a utilizar como primer ejemplo el partido de la fecha 24 donde River le ganó 3 a 1 a Boca de visitante. En este caso, River recibe la etiqueta 0 para el primer tipo de etiquetas ya que gano su encuentro mientras que Boca recibe la etiqueta 2 ya que perdió. River recibirá como etiqueta en diferencia de gol +2 mientras que Boca -2 pero ambos recibirán la misma etiqueta en cantidad de goles marcados en el partido que es 4.

Otro ejemplo es el encuentro de la fecha 9 donde Independiente y Rosario Central empataron sin goles. En este caso, ambos equipos reciben las mismas tres etiquetas. Un 1 para la etiqueta de resultado ya que empataron, mientras que para la diferencia de gol y goles marcados es un 0.

Notar que la primera etiqueta es un problema de clasificación mientras que las dos últimas etiquetas son un problema de regresión.

6.2. ¿Cómo esta compuesto el corpus?

Para armar estos modelos, **utilizamos 210 partidos** que corresponden a todos los encuentros en el torneo disputados por los equipos que tenemos el sentimiento de sus tweets. Recordar que este torneo tiene 30 fechas y que utilizamos a los equipos Boca, River, San Lorenzo, Racing, Independiente, Rosario Central, y Newell's.

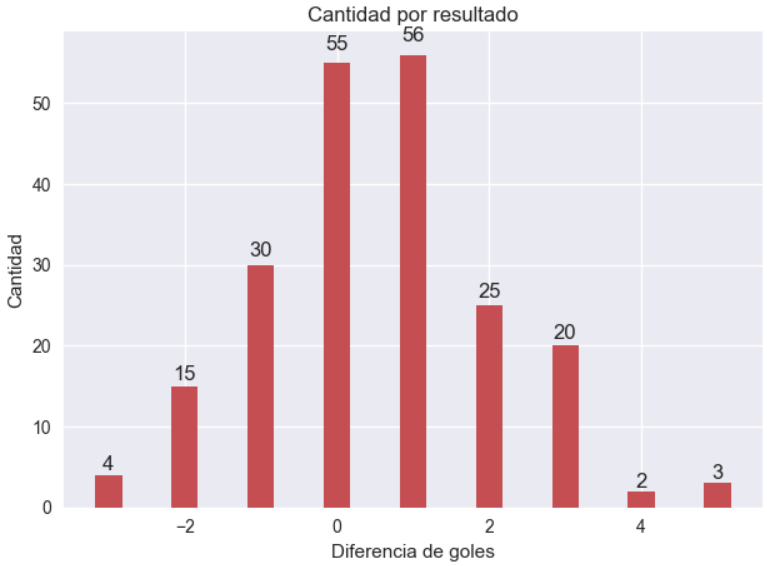

Fig. 6.1: Cantidad de partidos con una diferencia de gol

En Figure 6.1 podemos ver que hay mas partidos ganados por menos diferencia de goles y también hay mas partidos perdidos por menor diferencia de gol. Es decir, hay más partidos ganados por un gol que ganados por dos goles. También se puede ver que hay 3 partidos ganados por una diferencia de 5 goles mientras que 4 partidos perdidos por una diferencia de 3 goles. Notar que la cantidad de partidos con la etiqueta de diferencia de goles en 0, es decir terminó empatado, coincide con la cantidad de partidos empatados de Figure 4.3. Se puede pensar la etiqueta de diferencia de goles como ver la etiqueta de resultado del partido pero con más granularidad ya que un partido ganado tiene una diferencia de gol positiva mientras que un partido perdido una diferencia de gol negativa.

6.3. ¿Cómo construimos los modelos?

Para los predictores, usamos las siguientes ideas leídas de papers:

- **Utilizar atributos estadísticos y derivados de los tweets**
- **Contar las ocurrencias de ciertas palabras** como árbitro en la previa del partido ya que si se habla mucho del referí en la previa y el equipo juega de visitante, existe una posibilidad mayor a que el equipo pierda. También contar las ocurrencias de palabras como ganar, gana en la previa
- **Utilizar atributos estadísticos** como por ejemplo:
  - Porcentaje de victorias del equipo en condición de local para el local en la temporada actual mientras que el porcentaje de victorias del equipo visitante en condición de visitante en el mismo torneo

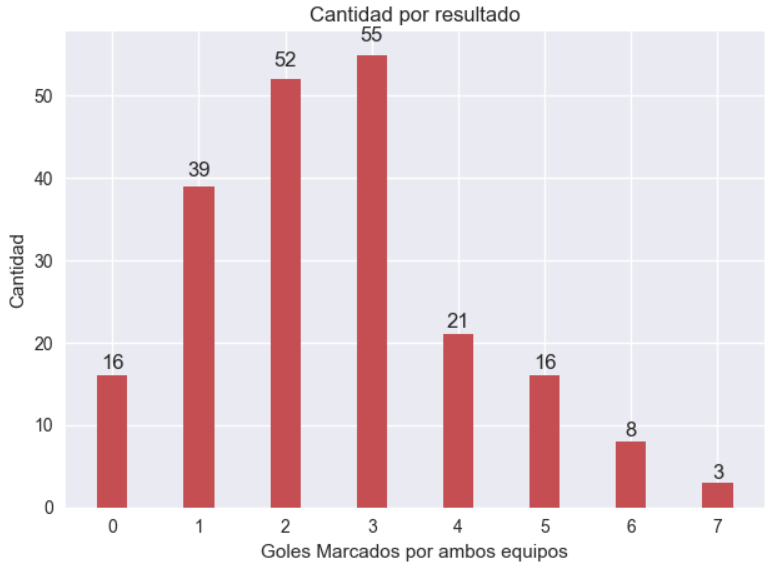

Fig. 6.2: Cantidad de partidos con una cantidad de goles marcados

- Promedio de goles anotados
- Promedio de la diferencia entre goles anotados y recibidos

6.3.1. Atributos estadísticos

En nuestra experimentación, usamos como atributos estadísticos:

1. Porcentaje de victorias del equipo en condición de local para el local en la temporada actual mientras que el porcentaje de victorias del equipo visitante en condición de visitante en el mismo torneo
2. Promedio de goles anotados y recibidos del local y del visitante
3. Promedio de la diferencia entre goles anotados y recibidos del local y del visitante
4. Porcentaje de rojas recibidas para el local y el visitante durante las fechas previas
5. Puntos y posición de ambos equipos en la condición que disputan el partido como en la general
6. Valor de ambos planteles en transfermarkt [1]
7. Cantidad de expulsados en contra del equipo en la fecha anterior
8. Cantidad de partidos ganados, perdidos y empatados en los últimos 5 partidos

[1] Transfermarkt

9. Condición de localía

10. Cantidad de partidos ganados, perdidos, empatados en la condición para ambos equipos. Es decir, cantidad de partidos ganados, perdidos, empatados de local para el equipo que hace de local mientras que partidos ganados, perdidos, empatados de visitante para el equipo que juega de visitante.

Una aclaración importante es que todos estos atributos salvo el atributo 6, utilizamos los datos hasta la fecha anterior. Por ejemplo, si queremos el promedio de goles marcados por Racing en condición de local en la fecha 7, usaremos de la primera fecha hasta la sexta. No usamos ni la fecha 7 ni el resto de las fechas.

6.3.2. Atributos de Twitter

En nuestra experimentación, usamos como tipos de atributos de Twitter:

1. Cantidad de tweets por minuto en la previa del partido actual

2. Cantidad de tweets por minuto en el post partido de la fecha anterior

3. Sentimiento por minuto utilizando las métricas Meaningcloud, Meaningcloud Normalizado y Facebook en la previa del partido actual

4. Sentimiento por minuto utilizando las métricas Meaningcloud, Meaningcloud Normalizado y Facebook en el post partido de la fecha anterior

5. Resumen de los puntos anteriores utilizando el mínimo, máximo, desvío estándar, promedio

6. Porcentaje de tweets que las palabras árbitro o referí figuran en los tweets desde 6 horas antes del comienzo del partido hasta su inicio

7. Similar al anterior pero utilizando solo la palabra jugadores

8. Porcentaje de tweets que figuran palabras relacionadas a fútbol en los tweets desde 6 horas antes del comienzo del partido hasta su inicio. Las palabras utilizadas son dale, ganó, huevo, gana, árbitro, vamo, referi, campeón, cancha, estadio, futbol, partido, pelota, gol, jugador, club, arquero, defensor, delantero, mediocampista

9. Porcentaje de tweets con los apodos de los equipos desde 6 horas antes del comienzo del partido hasta su inicio

10. Porcentaje de tweets que interactuaron con las cuentas semillas de los equipos desde 6 horas antes del comienzo del partido hasta su inicio

Una aclaración importante es que **los atributos de Twitter corresponden a un solo equipo del partido, que es el equipo que nos fijamos las etiquetas**. Por ejemplo, supongamos que miramos a River en el partido que le ganó a Boca por

3 a 1 de visitante en la fecha 24. Por ende, la primera etiqueta es un 0 ya que ganó, la segunda etiqueta un 2 y la tercera etiqueta un 4. En caso de utilizar atributos de Twitter, generamos los atributos con tweets de los hinchas de River. En caso de ver el partido desde el lado de Boca, la primera etiqueta es un 2 ya que perdió, la segunda etiqueta un -2 y la tercera etiqueta un 4 pero esta vez en caso de utilizar atributos de Twitter, generamos los atributos con tweets de los hinchas de Boca.

Decidimos ver los registros como partidos vistos desde un equipo en particular en lugar de ver si ganó el local, terminó en empate o ganó el visitante ya que al utilizar atributos de Twitter de ambos equipos sólo tenemos información de tan sólo 7 equipos reduciendo la cantidad de registros considerablemente ya que entre estos equipos, salvo los clásicos que se juegan dos veces en el torneo, sólo jugaron una vez entre sí generando un total de 24 encuentros.

Para los atributos 1 al 8, utilizamos para la previa del partido actual desde 6 horas antes hasta 1 horas antes cada una hora. Es decir, probamos utilizar el minuto a minuto desde 6 horas antes como también utilizando la última hora antes del comienzo del encuentro. Encambio, para el post partido de la fecha anterior utilizamos hasta 2 horas luego de la finalización del encuentro cada una hora. Por ejemplo, para el post partido de la fecha anterior vemos el minuto a minuto hasta una o dos horas luego de la finalización del partido.

Al utilizar atributos estadísticos, optamos siempre por utilizar todos los atributos estadísticos a diferencia de los atributos de Twitter que no siempre utilizamos todos.

6.3.3. Ejemplos del corpus

Como explicamos anteriormente, el corpus a entrenar puede tener distintos atributos y valores dependiendo de lo siguiente:

- ¿Cuáles tipos de atributos de Twitter usamos?
- La utilización o no de los atributos estadísticos
- ¿Cuánto tiempo miramos la previa del partido? ¿y el post partido de la fecha anterior?

Presentamos algunos ejemplos en Table 6.1,Table 6.2,Table 6.3 ,Table 6.4,Table 6.5,Table 6.6, y Table 6.7

Dos aclaraciones que nos parece importante destacar

1. En Table 6.2,y Table 6.3 utilizan los mismos atributos pero Table 6.2 utiliza tan solo 1 hora de previa mientras que Table 6.3 utiliza 2 horas de previa. Por ende, los atributos resumidos de Twitter tienen distintos valores. Utilizamos los mismos registros de partidos para que se pueda notar este cambio
2. En Table 6.1,Table 6.2,Table 6.3 ,Table 6.4,Table 6.5, y Table 6.7, las filas son distintos registros mientras que las columnas son los atributos. Sin embargo en Table 6.6 utilizamos otro formato para exhibir los registros con el fin de que sea legible ya que son 120 atributos

% victorias equipo	% victorias rival	Promedio goles marcados equipo	Promedio goles recibidos equipo	Promedio goles diferencias Equipo	Promedio goles marcados rival	Promedio goles recibidos rival	Promedio goles diferencias Rival	% Expulsados Equipo	Puntos equipo condición	Puntos equipo rival condición	Posición equipo condición	Posición equipo rival condición	Presupuesto rival	Presupuesto equipo	Expulsión en contra	Ganados por equipo	Empatados por equipo	Perdidos por equipo	Ganados por rival	Empatados por rival	Perdidos por rival	es local	Posición equipo	Posición rival	Puntos equipo	Puntos rival
1.000	0.000	2.000	0.000	2.000	0.000	1.000	-1.000	0.000	3	0	3	5	9	45	0	1	0	0	0	0	1	0	2	9	7	0
0.750	0.500	2.000	0.833	1.167	1.583	1.250	0.333	0.040	29	21	1	5	45	45	1	9	2	1	6	3	3	1	7	4	42	46
0.643	0.429	1.643	1.071	0.571	1.143	1.500	-0.357	0.107	28	19	4	7	13	45	1	9	1	4	6	1	7	1	5	2	49	54
0.286	0.125	0.714	0.857	-0.143	1.000	2.000	-1.000	0.133	8	7	11	9	13	46	1	2	2	3	1	4	3	1	6	17	25	14
0.600	0.600	1.400	0.600	0.800	0.800	0.200	0.600	0.000	11	11	4	1	46	45	0	3	2	0	3	2	0	1	4	6	18	16

Tab. 6.1: Utilizando solo estadisticas

% victorias equipo	% victorias rival	Promedio goles marcados equipo	Promedio goles recibidos equipo	Promedio goles diferencias Equipo	Promedio goles marcados rival	Promedio goles recibidos rival	Promedio goles diferencias Rival	% Expulsados Equipo	Puntos equipo condición	Puntos equipo rival condición	Posición equipo condición	Posición equipo rival condición	Presupuesto rival	Presupuesto equipo	Expulsión en contra	Ganados por equipo	Empatados por equipo	Perdidos por equipo	Ganados por rival	Empatados por rival	Perdidos por rival	es local	Posición equipo	Posición rival	Puntos equipo	Puntos rival	Menor valor	Mayor valor	Cantidad de 0s	Std	Promedio
0.455	0.100	1.200	0.700	0.500	0.400	0.700	-0.300	0.429	18	8	6	13	9	37	1	5	3	3	1	5	4	1	8	15	31	20	-4	3	22	1.252	0.197
0.750	0.250	1.750	0.583	1.167	0.750	1.667	-0.917	0.080	29	12	1	15	9	46	1	9	2	1	3	3	6	0	3	21	47	23	-5	9	12	2.588	0.918
0.625	0.000	1.875	1.375	0.500	0.571	2.143	-1.571	0.063	16	3	3	14	12	45	0	5	1	2	0	3	4	1	2	10	31	19	-2	7	20	1.608	0.803
0.714	0.071	2.357	0.786	1.571	0.714	1.571	-0.857	0.103	32	8	4	19	11	77	0	10	2	2	1	5	8	1	1	16	60	32	-2	21	2	5.069	6.246

Tab. 6.2: Utilizando el tipo de atributo de Twitter que resume la previa en la red social mirando una hora previa al partido y los atributos de estadisticas

% victorias equipo	% victorias rival	Promedio goles marcados equipo	Promedio goles recibidos equipo	Promedio goles diferencias Equipo	Promedio goles marcados rival	Promedio goles recibidos rival	Promedio goles diferencias Rival	% Expulsados Equipo	Puntos equipo condición	Puntos equipo rival condición	Posición equipo condición	Posición equipo rival condición	Presupuesto rival	Presupuesto equipo	Expulsión en contra	Ganados por equipo	Empatados por equipo	Perdidos por equipo	Ganados por rival	Empatados por rival	Perdidos por rival	es local	Posición equipo	Posición rival	Puntos equipo	Puntos rival	Menor valor	Mayor valor	Cantidad de 0s	Std	Promedio
0.455	0.100	1.200	0.700	0.500	0.400	0.700	-0.300	0.429	18	8	6	13	9	37	1	5	3	3	1	5	4	1	8	15	31	20	-4	5	43	1.414	0.314
0.750	0.250	1.750	0.583	1.167	0.750	1.667	-0.917	0.080	29	12	1	15	9	46	1	9	2	1	3	3	6	0	3	21	47	23	-5	9	21	2.511	0.612
0.625	0.000	1.875	1.375	0.500	0.571	2.143	-1.571	0.063	16	3	3	14	12	45	0	5	1	2	0	3	4	1	2	10	31	19	-4	7	35	1.528	0.752
0.714	0.071	2.357	0.786	1.571	0.714	1.571	-0.857	0.103	32	8	4	19	11	77	0	10	2	2	1	5	8	1	1	16	60	32	-2	21	7	4.475	5.149

Tab. 6.3: Utilizando el tipo de atributo de Twitter que resume la previa en la red social mirando dos horas previas al partido y los atributos de estadisticas

60 minutos antes	59 minutos antes	58 minutos antes	57 minutos antes	56 minutos antes	55 minutos antes	54 minutos antes	53 minutos antes	52 minutos antes	51 minutos antes	50 minutos antes	49 minutos antes	48 minutos antes	47 minutos antes	46 minutos antes	45 minutos antes	44 minutos antes	43 minutos antes	42 minutos antes	41 minutos antes	40 minutos antes	39 minutos antes	38 minutos antes	37 minutos antes	36 minutos antes	35 minutos antes	34 minutos antes	33 minutos antes	32 minutos antes	31 minutos antes	30 minutos antes	29 minutos antes	28 minutos antes	27 minutos antes	26 minutos antes	25 minutos antes	24 minutos antes	23 minutos antes	22 minutos antes	21 minutos antes	20 minutos antes	19 minutos antes	18 minutos antes	17 minutos antes	16 minutos antes	15 minutos antes	14 minutos antes	13 minutos antes	12 minutos antes	11 minutos antes	10 minutos antes	9 minutos antes	8 minutos antes	7 minutos antes	6 minutos antes	5 minutos antes	4 minutos antes	3 minutos antes	2 minutos antes	1 minuto antes
6	4	6	3	2	-1	4	3	3	1	3	5	-1	-1	1	-2	-1	1	5	2	-1	2	2	3	1	-1	-1	-1	4	0	-1	1	-1	4	3	0	0	1	2	2	1	2	4	-4	-2	1	3	-3	1	2	2	4	2	1	6	3	2	1	1	-1
1	2	0	2	1	1	1	3	0	1	2	0	0	0	0	0	1	-1	0	0	3	2	4	0	1	2	2	4	0	0	0	0	0	1	1	0	0	1	1	2	2	4	5	2	1	0	2	1	0	1	1	1	4	1	0	4	5	5	6	3
-1	-1	1	0	0	0	1	1	0	0	0	1	0	0	0	2	0	1	0	0	1	0	0	3	-1	-1	-1	-2	0	0	-1	0	1	3	0	1	1	2	1	2	0	1	0	1	0	1	2	0	-1	0	1	0	1	1	0	1	1	0	0	-3
0	0	0	-1	4	1	0	2	1	0	0	0	0	1	2	0	0	0	1	0	0	0	0	2	3	2	0	-1	0	0	0	-1	0	2	0	1	2	1	2	2	-2	-1	0	2	1	0	1	2	1	0	0	1	-1	1	0	-1	1	0	-2	1

Tab. 6.4: Utilizando el tipo de atributo de Twitter que mide el sentimiento minuto a minuto con la métrica Meaning Cloud en la red social mirando una hora previa al partido

	Facebook			MeaningCloud		
Minutos antes	Ejemplo 1	Ejemplo 2	Ejemplo 3	Ejemplo 1	Ejemplo 2	Ejemplo 3
1	-0.007	0.026	-0.070	2	0	-2
2	0.025	0.079	-0.021	1	2	-3
3	0.011	-0.015	-0.100	2	0	-3
4	0.047	0.131	0.056	5	3	2
5	0.012	0.128	-0.060	3	2	-3
6	0.029	0.004	0.023	1	0	0
7	0.048	0.000	-0.021	4	0	-2
8	-0.003	0.071	-0.003	0	1	1
9	0.018	0.273	0.002	0	1	-2
10	0.029	0.074	0.016	1	4	1
11	-0.001	0.026	-0.019	1	4	-4
12	0.044	0.005	0.049	2	0	4
13	0.054	0.000	-0.001	3	0	0
14	0.077	0.042	0.004	1	2	0
15	-0.006	0.040	0.013	-1	3	1
16	-0.067	-0.050	-0.009	-1	-2	-2
17	0.008	0.030	-0.015	1	2	2
18	0.019	0.049	0.045	1	2	4
19	-0.036	-0.033	-0.024	-1	-1	-2
20	0.000	0.014	-0.019	0	0	0
21	-0.010	0.013	-0.031	0	1	-2
22	0.077	0.000	0.018	5	0	0
23	0.164	0.042	0.250	2	2	1
24	0.000	-0.125	-0.044	-1	-1	-3
25	0.084	-0.060	0.070	3	-2	4
26	0.052	0.013	-0.023	8	1	-3
27	0.010	0.012	0.000	0	-1	0
28	-0.071	0.019	-0.083	2	0	-1
29	0.083	0.000	0.031	5	0	1
30	0.050	0.042	-0.024	5	1	-3
31	0.020	0.028	0.028	2	1	0
32	0.008	0.077	-0.068	1	2	-2
33	-0.049	0.042	-0.028	-4	1	-2
34	0.096	0.032	-0.037	3	1	-2
35	0.022	0.101	0.082	2	2	4
36	0.051	0.015	-0.053	2	0	-2
37	0.069	0.131	0.031	3	2	2
38	0.123	0.000	0.012	5	0	0
39	0.044	0.061	-0.054	1	2	-2
40	0.011	0.010	0.030	1	0	3
41	0.015	0.000	-0.009	0	0	0
42	0.015	0.014	-0.020	4	-1	-1
43	0.050	0.013	0.039	3	-1	5
44	0.106	-0.032	0.000	6	-1	0
45	-0.030	0.026	0.000	-2	2	0
46	0.063	0.029	-0.050	3	0	-2
47	0.025	-0.016	-0.026	0	0	-2
48	-0.009	0.023	0.031	-1	1	2
49	0.030	0.044	-0.041	2	2	-1
50	0.024	0.019	0.022	2	1	3
51	0.026	0.090	-0.022	5	4	3
52	-0.005	0.000	0.045	0	0	5
53	0.056	0.000	0.054	6	0	5
54	0.019	0.023	-0.027	3	1	-7
55	0.067	0.143	-0.003	7	1	-1
56	0.038	0.043	0.041	5	1	0
57	0.069	0.028	0.012	13	2	0
58	0.057	0.000	0.028	8	0	0
59	0.043	0.061	0.039	6	2	1
60	0.031	0.129	-0.072	-1	4	-1

Tab. 6.5: Utilizando el tipo de atributo de Twitter que mide el sentimiento minuto a minuto con las métricas Meaning Cloud y Facebook en la red social mirando una hora previa al partido

% victorias equipo	% victorias rival	Promedio goles marcados equipo	Promedio goles recibidos equipo	Promedio goles diferencias Equipo	Promedio goles marcados rival	Promedio goles recibidos rival	Promedio goles diferencias Rival	% Expulsados Equipo	Puntos equipo condición	Puntos equipo rival condición	Posición equipo condición	Posición equipo rival condición	Presupuesto rival	Presupuesto equipo	Expulsión en contra	Ganados por equipo	Empatados por equipo	Perdidos por equipo	Ganados por rival	Empatados por rival	Perdidos por rival	es local	Posición equipo	Posición rival	Puntos equipo	Puntos rival	Menor valor Previa Twitter	Mayor valor Previa Twitter	Cantidad de 0s Previa Twitter	Std Previa Twitter	Promedio Previa Twitter	Menor valor Post Twitter	Mayor valor Post Twitter	Cantidad de 0s Post Twitter	Std Post Twitter	Promedio Post Twitter
0.429	0.071	1.000	0.714	0.286	0.500	0.857	-0.357	0.143	22	10	5	18	22	27	1	6	4	4	1	7	6	0	5	22	49	22	-4	6	4	2.191	1.410	-5	6	11	2.255	1.413
0.750	0.333	2.125	0.750	1.375	0.889	1.000	-0.111	0.118	19	10	2	8	8	27	1	6	1	1	3	1	5	1	3	12	32	21	-3	4	24	1.378	0.344	-5	3	49	1.300	0.380
0.500	0.500	1.000	0.500	0.500	1.000	0.500	0.500	0.000	4	4	3	2	13	46	0	1	1	0	1	1	0	0	3	6	8	5	-4	3	15	1.294	0.361	-6	12	40	2.220	-0.116
0.500	0.286	1.875	1.000	0.875	0.857	1.571	-0.714	0.188	15	7	2	12	13	77	1	4	3	1	2	1	4	0	1	14	34	15	-3	7	9	2.114	1.082	-3	16	25	3.193	1.793

Tab. 6.6: Utilizando el tipo de atributo de Twitter que resume lo sucedido en la red social mirando una hora previa al partido y una hora post partido de la fecha anterior y los atributos de estadisticas

6.4. Experimentación para obtener los mejores modelos

Por un tema de implementación para poder reproducir los resultados obtenidos, utilizamos como semilla el número 22. Es decir, al crear los modelos mencionados seteamos como random_state al número 22.

Utilizamos un 0.7 % de los registros para el set de entrenamiento y el 0.3 % restante para el set de test. Entrenamos el modelo utilizando el set de entrenamiento y luego nos fijamos la exactitud en el set de test. Para entrenar el modelo, utilizamos la función GridSearchCV de sklearn que entrena un modelo utilizando la técnica de cross-validation y además busca los mejores parámetros que configuran al modelo. En particular, utilizamos 5 folds.

Separamos la experimentación en dos etapas.

6.4.1. Primera etapa

Entrenamos los modelos utilizando o no los primeros 8 atributos obtenidos de Twitter y los estadísticos mirando los atributos de Twitter que utilizan información de la previa desde 6 horas antes del partido hasta el comienzo cada uno hora y mirando los que utilizan información post partido desde que finaliza el partido hasta dos horas luego mirando también cada una hora.

Los modelos que entrenamos en la primera etapa son

- Regression logística
- SVM con kernels lineal o sigmoid o rbf
- Random Forest
- Adaptive Boosting, mejor conocido como AdaBoost
- Potenciación del gradiente, mejor conocido como GradientBoosting
- ExtraTrees

	Ejemplo 1	Ejemplo 2	Ejemplo 3
% victorias equipo	0.600	0.462	0.500
% victorias rival	0.200	0.462	0.750
Promedio goles marcados equipo	1.800	1.000	1.750
Promedio goles recibidos equipo	1.200	0.615	0.500
Promedio goles diferencias Equipo	0.600	0.385	1.250
Promedio goles marcados rival	1.400	0.923	1.000
Promedio goles recibidos rival	1.000	0.846	0.250
Promedio goles diferencias Rival	0.400	0.077	0.750
% Expulsados Equipo	0.100	0.115	0.500
Puntos equipo condición	10	22	7
Puntos equipo rival condición	6	21	10
Posición equipo condición	5	5	5
Posición equipo rival condición	5	9	1
Presupuesto rival	77	11	46
Presupuesto equipo	45	27	37
Expulsión en contra	0	1	1
Ganados por equipo	3	6	2
Empatados por equipo	1	4	1
Perdidos por equipo	1	3	1
Ganados por rival	1	6	3
Empatados por rival	3	3	1
Perdidos por rival	1	4	0
es local	1	0	1
Posición equipo	2	3	9
Posición rival	3	15	3
Puntos equipo	21	48	9
Puntos rival	19	29	15
60 minutos antes	-1	-2	2
59 minutos antes	0	-3	0
58 minutos antes	3	2	0
57 minutos antes	3	-1	-1
56 minutos antes	-1	0	0
55 minutos antes	0	2	-1
54 minutos antes	0	1	1
53 minutos antes	1	-7	-1
52 minutos antes	-1	1	1
51 minutos antes	0	-1	3
50 minutos antes	-2	3	0
49 minutos antes	2	0	-2
48 minutos antes	1	5	1
47 minutos antes	2	3	0
46 minutos antes	2	3	-2
45 minutos antes	1	1	1
44 minutos antes	-2	4	-1
43 minutos antes	11	2	1
42 minutos antes	24	2	-1
41 minutos antes	20	0	1
40 minutos antes	23	4	0
39 minutos antes	18	3	1
38 minutos antes	15	-3	0
37 minutos antes	17	0	0
36 minutos antes	11	-1	0
35 minutos antes	23	6	0
34 minutos antes	12	-1	0
33 minutos antes	9	1	-2
32 minutos antes	13	0	0
31 minutos antes	8	4	0
30 minutos antes	8	-10	1
29 minutos antes	2	3	0
28 minutos antes	10	5	0
27 minutos antes	9	0	0
26 minutos antes	5	-1	-1
25 minutos antes	8	-2	1
24 minutos antes	5	2	0
23 minutos antes	6	-3	0
22 minutos antes	6	2	0
21 minutos antes	5	0	0
20 minutos antes	0	-2	-1
19 minutos antes	5	2	-1
18 minutos antes	6	-2	-1
17 minutos antes	8	0	0
16 minutos antes	2	4	-1
15 minutos antes	4	0	1
14 minutos antes	1	2	2
13 minutos antes	2	7	0
12 minutos antes	1	5	0
11 minutos antes	8	0	-2
10 minutos antes	0	3	-1
9 minutos antes	12	-1	0
8 minutos antes	4	2	0
7 minutos antes	4	-2	1
6 minutos antes	4	-2	0
5 minutos antes	6	3	1
4 minutos antes	8	-3	0
3 minutos antes	2	1	-2
2 minutos antes	6	0	0
1 minuto antes	3	-1	0
Menor valor Previa Twitter	-5	-10	-2
Mayor valor Previa Twitter	24	7	3
Cantidad de 0s Previa Twitter	6	12	29
Std Previa Twitter	6.49	2.91	1.02
Promedio Previa Twitter	5.85	0.66	-0.07

Tab. 6.7: Utilizando los tipos de atributo de Twitter que miden el sentimiento minuto a minuto y el resumen de estos valores con la métrica Meaning Cloud en la red social mirando una hora previa al partido y los atributos de las estadísticas

Notar que en esta etapa lo que hacemos es probar todas las posibles combinaciones de seleccionar los primeros ocho atributos de Twitter juntando o no con todos los atributos estadísticos. Al ser 8 posibles atributos de Twitter, obtenemos 255 posibilidades pero al mirar o no los atributos estadísticos, obtenemos 510.

Al entrenar con 8 modelos distintos, obtenemos 4.080 posibilidades distintas fijados en un tiempo de mirar la previa del partido actual y el post partido del partido anterior.

No solo probamos todas las posibilidades de seleccionar los atributos de la red social, sino que también desde y hasta cuando miramos la información de Twitter.

Notar que no es lo mismo utilizar los atributos de la red social utilizando información de tal desde 1 hora que desde 2 horas antes del comienzo del partido. Por ejemplo, en la primera opción al utilizar el primer atributo de Twitter, voy a obtener 60 valores que representan la cantidad de tweets minuto a minuto desde una hora antes hasta el comienzo pero en la segunda opción al utilizar el mismo atributo, voy a obtener 120 valores que representan el minuto a minuto desde 2 horas antes del inicio del encuentro.

Por ende, como miramos desde 6 horas antes del comienzo del partido para la previa del partido actual tenemos 6 posibilidades mientras que para el post partido del partido anterior miramos hasta 2 horas luego de la finalización tenemos 2 posibilidades, generando un total de 12 posibilidades en referencia de los tiempos que mira.

6.4.2. Segunda etapa

Utilizamos los modelos que usan información de Twitter que tienen mayor exactitud que los modelos que solo utilizan atributos estadísticos. En esta etapa, probamos si al agregar los atributos del 6 al 10 de Twitter obtenemos una mayor exactitud.

6.4.3. Otro modelo para predecir el resultado del partido

También intentamos utilizar el mejor modelo de diferencia de gol para construir un modelo que trate de predecir el resultado del partido, donde simplemente se cambia el valor de la etiqueta devuelta por el modelo de diferencia de gol.

- Si la etiqueta devuelta por el modelo de diferencia de goles tiene un valor negativo, se le asigna como etiqueta un 2 ya que significa que perdió porque la diferencia de goles es negativa

- Si la misma etiqueta tiene un valor positivo, se le asigna como etiqueta un 0 ya que significa que ganó porque la diferencia de goles es positiva

- Si la etiqueta del primer modelo tiene un valor igual a 0, se la asigna como etiqueta un 1 ya que empató al no tener diferencia de goles

Llamamos a este modelo convertido.

Exactitud	Resultado del partido	Diferencia de gol	Goles marcados
Solo atributos estadisticos	0.507936507937	0.349206349206	0.349206349206
Primera etapa	0.634920634921	0.412698412698	0.460317460317
Segunda etapa	0.634920634921	0.412698412698	0.428571428571

Tab. 6.8: Exactitud obtenida con los mejores modelos

6.5. Resultados

En Table 6.8 se puede ver que **se obtienen mejores modelos al utilizar información de Twitter con los datos estadísticos**. Sin embargo, al intentar agregar más atributos de la red social no logramos mejorar la exactitud en ninguno de los tres casos.

El modelo convertido logro obtener una exactitud de 0.52380952381 siendo peor que la obtenida con el primer modelo construido para predecir el resultado del partido que utiliza como modelo un SVM con kernel lineal utilizando información de Twitter y datos estadísticos. Particularmente, utiliza las 5 horas de la previa en Twitter del partido utilizando el segundo y sexto tipos de atributos de la red social.

Para tratar de predecir la cantidad de goles por partido, el modelo que obtuvo la mejor exactitud utiliza un modelo Adaptive Boosting donde solo utiliza los datos de la red social.

Sin embargo, tan sólo utiliza de Twitter las cantidades de Tweets desde 2 horas antes del comienzo del partido de la fecha actual y hasta 1 hora luego de la finalización del partido de la fecha anterior. Además, las utiliza en formato del segundo tipo de atributos de la red social. Es decir, resumidos.

Por último, en Table 6.9 se encuentra las configuraciones de los ocho modelos que obtienen la mejor exactitud al predecir la diferencia de gol.

cantidad horas en la previa	cantidad horas en post partido de fecha anterior	usa estadisticas	usa datos de Twitter en el post partido de fecha anterior	usa datos de Twitter en la previa del partido de la fecha actual	usa datos resumidos de Twitter en la previa del partido de la fecha actual	usa datos resumidos de Twitter en el post partido de fecha anterior	usa metrica Facebook	@usa metrica MeaningCloud	@usa metrica MeaningCloud Normalizada	@usa Cantidades
1	2	✓		✓		✓	✓	✓		
2	1		✓	✓	✓		✓			
2	1		✓	✓			✓			
2	1		✓	✓			✓			
2	2			✓	✓	✓		✓	✓	✓
3	1		✓	✓	✓		✓		✓	
4	2		✓	✓	✓		✓			
6	2	✓	✓	✓				✓	✓	

Tab. 6.9: Configuraciones de los modelos que obtienen la mejor exactitud al predecir la diferencia de gol

7. CONCLUSIONES

En este trabajo por un lado estudiamos el comportamiento del hincha Argentino en la red social Twitter. Pensamos que pese a ser preguntas triviales, nos sirven no solo para validar estas preguntas sino también para validar el método propuesto y construir el predictor de resultados. Sin embargo para lograr eso tuvimos que primero encontrar una forma de poder obtener hinchas en la red social Twitter de los equipos a analizar. Para realizar esto utilizamos a los seguidores de las cuentas semillas que cumplan 4 condiciones en el perfil de la red social que analizan su biografía, a quienes siguen y sus tweets. Para conseguir esta información de la red social, implementamos una herramienta que divide la totalidad de Twitter apps en una cantidad fija de threads donde cada uno de los hilos se encarga de administrar los tickets de sus twitter apps consumiendo primero todos los tickets de una Twitter app. Una vez conseguido los hinchas de los clubes con sus respectivos tweets, le realizamos análisis de sentimiento a tales utilizando el servicio de MeaningCloud y mediante 3 métricas creadas por nosotros estudiamos el comportamiento de los hinchas en Twitter.

Por otro lado entrenamos modelos para inferir el resultado del partido, la cantidad de goles marcados, y la diferencia de gol. Para realizar tal tarea utilizamos tanto estadísticas de los equipos como información obtenida de la red social. Podemos concluir lo siguiente

- Podemos destacar como resultados del comportamiento que no solo el hincha twittea más al ganar que al perder sino que también tiene mejor humor, el cual dura 30 horas aproximadamente por el efecto del resultado

- Obtuvimos una exactitud de 0.99 % al etiquetar cuentas de Twitter validado por colaboradores en un experimento

- Cuanto más Twitter apps se tenga para repartir entre los threads, mayor será la velocidad de descargar información de Twitter

- Las métricas creadas sirven para analizar nuestro problema

- La herramienta de análisis de sentimiento configurada devuelve resultados correctos ya que sin configurar un simple tweet como el festejo de un gol, no lo etiquetaba correctamente

- Utilizando información de Twitter se obtiene mejor exactitud que al utilizar solamente estadísticas de los equipos

- Se obtiene una exactitud de 0.63 % al predecir resultados de partidos. Pese a que creemos que aún hay mucho por trabajar en estos modelos, creemos que es un buen punto de partido

7.1. Debilidades

Pese a que intentamos realizar lo mejor de nuestro lado, sabemos que nuestro trabajo tiene las siguientes debilidades

- Pese a que el servicio de MeaningCloud nos indique si el fragmento de texto es irónico o no, no analizamos la ironía de los tweets en nuestro análisis de sentimiento
- No solo que tenemos pocos equipos con información de Twitter para utilizar en los predictores, sino que de los equipos que tenemos información es tan solo de un torneo. Esto impacta en la cantidad de registros para entrenar el modelo ya que al tener más partidos o equipos, crece la muestra
- No tener los hinchas etiquetados desde el comienzo del torneo nos perjudico en obtener todos los tweets de los hinchas que escribieron más de 3.200 tweets ya que la API de Twitter permite obtener los últimos 3.200. Esto impactó en que terminamos teniendo más tweets en las últimas fechas que en las primeras

8. TRABAJO FUTURO

Pensamos que sería interesante seguir este trabajo analizando los siguientes puntos:

- Analizar la ironia. Algunas preguntas que se nos ocurren es si existe algún equipo que sus hinchas jueguen más con la ironía, si esta relación tiene que ver con su ubicación en la tabla o espacialmente en el país. También analizar cómo impacta la ironía en el análisis de sentimiento, siendo la pregunta a hacer si el análisis de sentimiento es capaz de poder detectar la ironía?
- Conseguir nuevos equipos con hinchas reconocidos y más hinchas de los equipos actuales para agregar más información de Twitter en los predictores. Para tener mayor cantidad de hinchas en los equipos que ya tenemos hinchas buscar en todas las cuentas semillas, los seguidores que cumplan las 4 condiciones
- Al conseguir equipos que no sean de provincia de Buenos Aires con hinchas etiquetados en Twitter, comparar los equipos de esta provincia contra los del resto. Preguntas que se nos ocurren son ¿quién alienta más? ¿Qué sucede cuando un equipo de otra provincia juega en Buenos Aires?
- Analizar los equipos que juegan copas internacionales. ¿Tienen más actividad en la red social cuando juegan estos certámenes? ¿Tienen el mismo humor al perder o ganar un partido en una copa internacional que en torneo local?
- Analizar el resultado de utilizar nuestro modelo para inferir resultados mirando la ganancia o pérdida a obtener al apostar en casas de apuestas
- Mejorar los predictores ya que al crecer la muestra por tener tanto más equipos y partidos, se pueden obtener nuevas conclusiones

9. ANEXO DE RESULTADOS

9.1. Análisis de sentimiento por localia

En esta sección nos planteamos ver **la relación entre la localía y el resultado final del partido en el humor del hincha**.

9.1.1. ¿Se disfruta más una victoria jugando en condición de local o de visitante?

Nuestra hipótesis es que ganar un partido de visitante genera una mayor euforia y felicidad que ganar un partido de local.

En Figure 9.1 se puede ver la evolución del humor del hincha con la métrica Meaning cloud normalizada con los partidos ganados utilizando 396.102 tweets.

El pvalue de comparar los valor de la métrica meaning cloud normalizada en local contra visitante en partidos ganados es 5.065e−14.
Lo que se puede observar en Figure 9.1 es que se obtiene mejor valor de la métrica cuando el equipo gana de local que de visitante, contradiciendo nuestra hipótesis.
Pensábamos que se debía a que alguna condición de localía no llegó a estabilizar el valor de la métrica pero a pesar de tener 93 partidos ganados de local con 216.508 tweets, tenemos 71 partidos ganados de visitante con 179.594 tweets con el valor de la métrica de sentimiento. Pese a no tener una cantidad similar de tweets para partidos ganados en ambas condiciones de localia, creemos que con la cantidad de tweets que se tienen en condición de visitante se alcanza a estabilizar la métrica.

9.1.2. ¿Se sufre más una derrota jugando en condición de local o de visitante?

Nuestra hipótesis es que perder un partido de local genera peor humor en el hincha que una derrota de visitante ya que de local el equipo tiene que ganar.

En Figure 9.2 se puede ver la evolución del humor del hincha con los encuentros que terminan en derrota utilizando 254.819 tweets.

Valor de la media	Local	Visitante
Media	0.328	0.275

Tab. 9.1: Valor de la media para la metrica Meaningcloud Normalizado filtrado por localia para partidos ganados

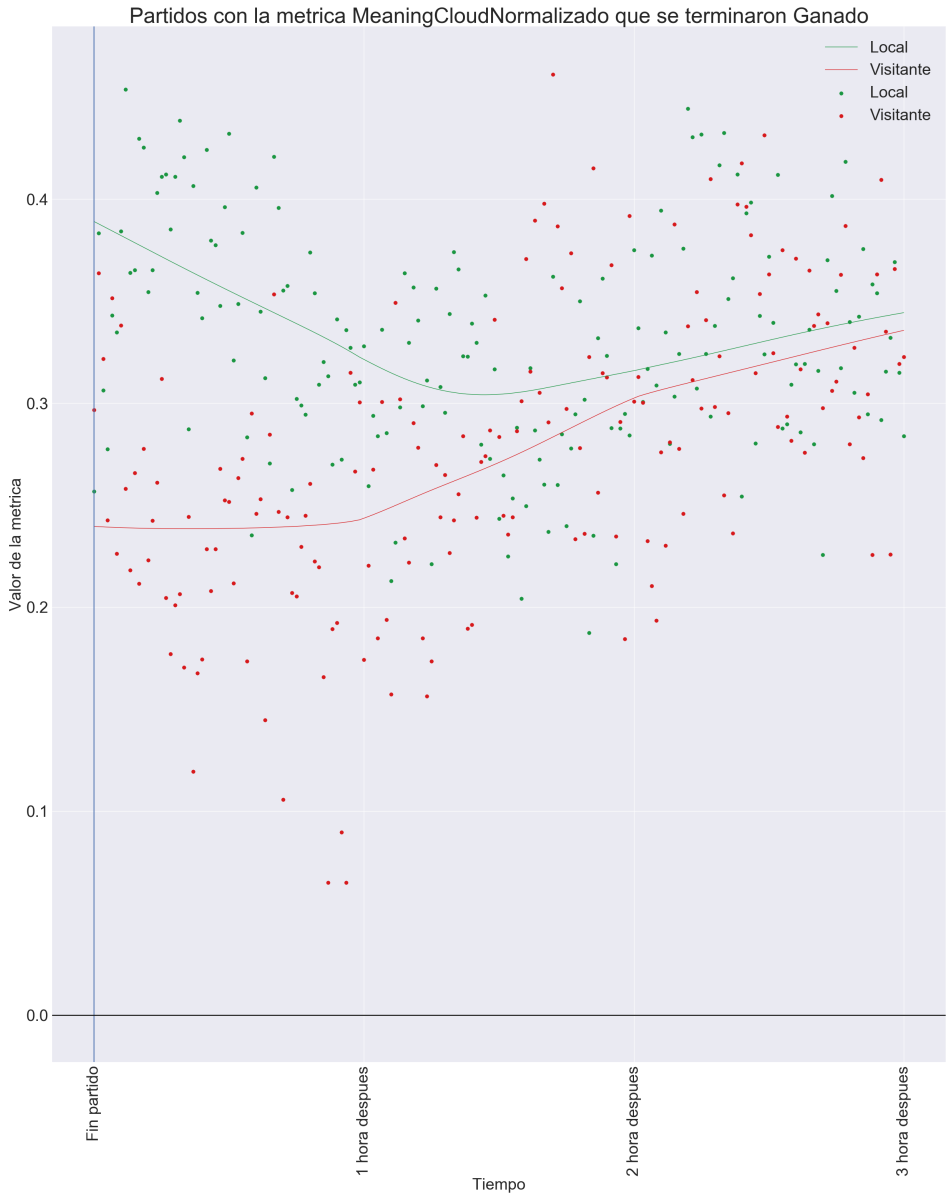

Fig. 9.1: Evolución del humor del hincha post partido con la métrica Meaning cloud normalizada con los partidos ganados filtrado por localía

Valor de la media	Local	Visitante
Media	0.016	-0.063

Tab. 9.2: Valor de la media para la metrica Meaningcloud Normalizado filtrado por localia para partidos perdidos

El pvalue de comparar los valor de la métrica meaning cloud normalizada en local contra visitante en partidos perdidos es 8.112e−17.

En Figure 9.2 se puede observar que se obtiene peores valores de la métrica al perder de visitante que de local, rechazando nuestra hipótesis. Pensábamos que se debía a que alguna condición de localía no llegó a estabilizar el valor de la métrica pero a pesar de tener 37 partidos perdidos de local y 61 de visitante tenemos 98.811 y 156.008 tweets respectivamente con el valor de la métrica de sentimiento. Llegamos al mismo razonamiento que con los partidos ganados.

Los resultados anteriormente mostrados para la métrica Meaning cloud normalizada son similares a los obtenidos con las otras dos métricas

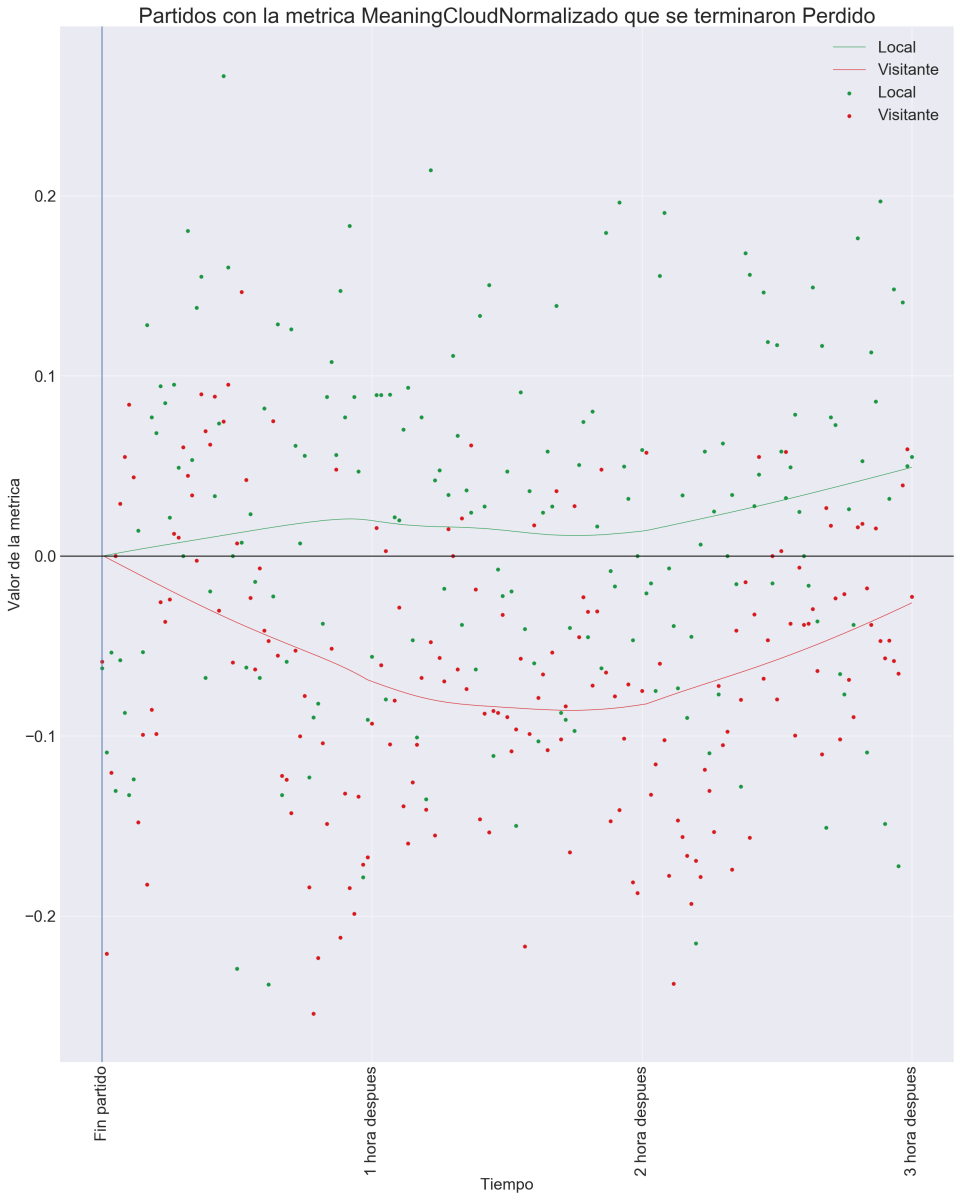

Fig. 9.2: Evolución del humor del hincha post partido con la métrica Meaning cloud normalizada con los partidos perdidos filtrado por localía

Resultado	Difíciles			Ganables		
	Todos	Local	Visitante	Todos	Local	Visitante
Ganados	23	14	9	30	19	11
Empatados	12	8	4	17	10	7
Perdidos	26	10	16	5	3	2

Tab. 9.3: Cantidad de partidos por etiqueta

9.2. Análisis de sentimiento de partidos por dificultad

En esta sección comparamos partidos separados por su dificultad, las cuales son

- Difícil: Se enfrenta contra un equipo que se encuentra entre las primeras posiciones o es un equipo de los denominados grandes
- Fácil: El rival está en las últimas posiciones de la tabla y no es un equipo de los denominados grandes

Etiquetamos los partidos a mano pero no etiquetamos todos los partidos ya que no todos cumplen estas condiciones. En la tabla 9.3 se puede ver por cada etiqueta cuantos partidos se jugaron de local y visitante que terminaron ganados o empatados o perdidos.

9.2.1. ¿Hay una diferencia en el humor del hincha dependiendo de la localía al jugar un partido etiquetado como difícil? ¿Y al jugar un partido etiquetado como fácil?

La primera pregunta que nos hicimos en torno a la dificultad es si hay una diferencia en el humor del hincha dependiendo de la localía. Es decir,¿ los hinchas se comportan igual en partidos etiquetados como difíciles de local que de visitante? ¿Y cómo se comportan en los partidos etiquetados como fáciles? Nuestra hipótesis para ambas preguntas es que el hincha se encuentra de mejor humor de local que de visitante ya que tiene más posibilidad de ganar.

Lo que podemos ver en Figure 9.4 es que se cumplen nuestra hipótesis mientras que en Figure 9.3 es muy similar el humor del hincha salvo la primera hora post partido donde se ve un mejor humor en los hinchas que su equipo jugó de local.

Para Figure 9.3 donde se analizan los partidos etiquetados como difíciles se utilizaron 607.824 tweets donde 302.035 corresponden a los partidos disputados de local mientras que 305.789 a los de visitante. Por otro lado, para Figure 9.4 donde se analizan los partidos etiquetados como ganables utilizamos 180.169 tweets donde 100.301 corresponden a partidos disputados en condición de local mientras que el resto en condición de visitante.

Fig. 9.3: Minuto a minuto desde el comienzo del partido hasta 3 horas después de la finalizacion filtrado por localia en partidos etiquetados como difíciles

Valor de la media	Durante el partido	Post partido
Local	0.171	0.155
Visitante	0.156	0.137

Tab. 9.4: Valor de la media por periodo filtrado por localia utilizando la métrica MeaningCloud normalizada en partidos difíciles

P-value	Durante el partido	Post partido
Local vs Visitante	0.045	0.002

Tab. 9.5: P-value por periodo comparando por localia utilizando la métrica MeaningCloud normalizada en partidos difíciles

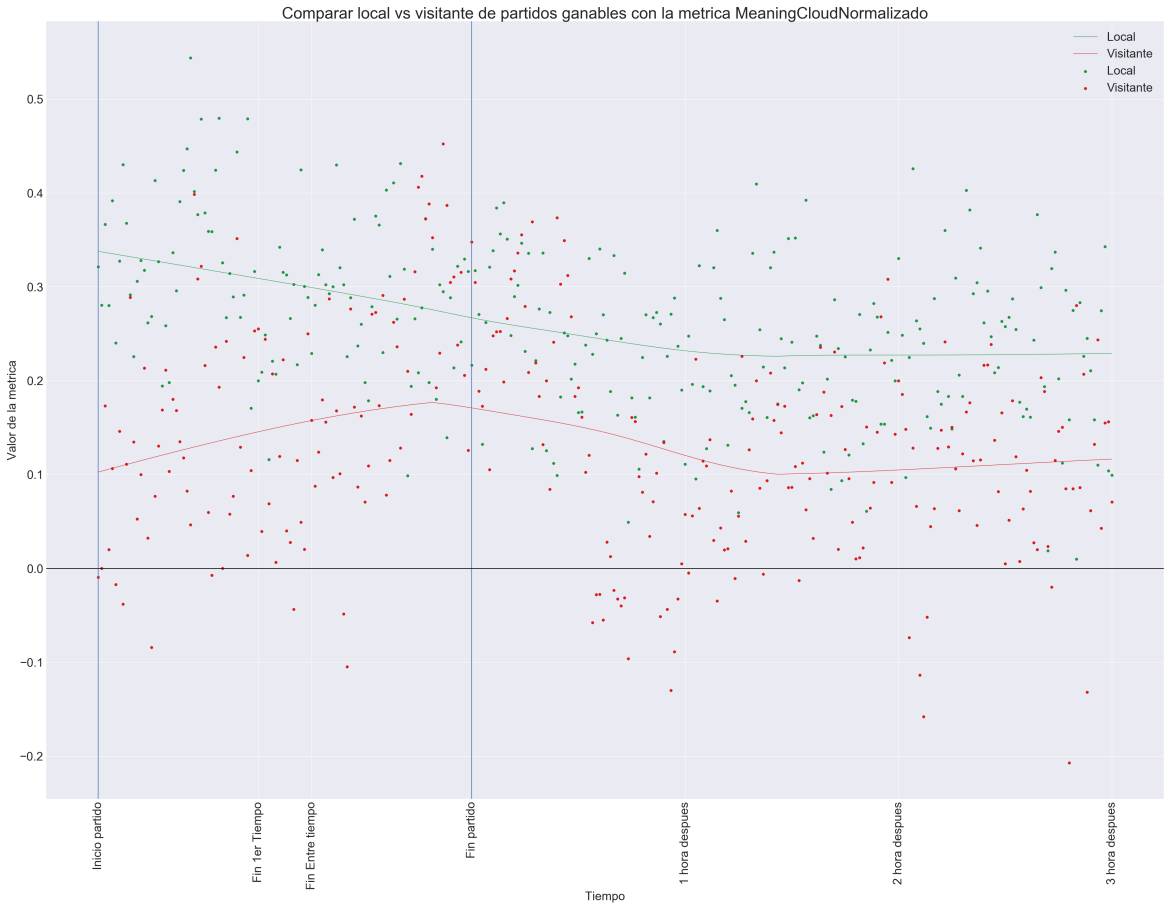

Fig. 9.4: Minuto a minuto desde el comienzo del partido hasta 3 horas después de la finalizacion filtrado por localia en partidos etiquetados como fáciles

Valor de la media	Durante el partido	Post partido
Local	0.304	0.231
Visitante	0.164	0.114

Tab. 9.6: Valor de la media por periodo filtrado por localia utilizando la métrica MeaningCloud normalizada en partidos fáciles

P-value	Durante el partido	Post partido
Local vs Visitante	1.512e−18	1.288e−25

Tab. 9.7: P-value por periodo comparando por localia utilizando la métrica MeaningCloud normalizada en partidos fáciles

	Ganables	Dificiles
Media	0.229	0.177

Tab. 9.8: Media comparando por dificultad utilizando la metrica MeaningCloud normalizada en condiccion de local en la previa

Pensábamos que el resultado obtenido en la primera hora al finalizar los partidos con mayor exigencia se debe a que de visitante se perdió más partidos de los que se ganó, mientras que de local se ganó más. Sin embargo, en condición de visitante se ganó tan solo 9 partidos generando 198.531 tweets mientras que se perdió 16 generando 182.368 tweets y se empató 4 generando 34.919 tweets y en condición de local se ganó 14 encuentros generando 194.008 tweets, se perdieron 10 generando 186.978 tweets y se empataron 8 generando 35.686 tweets.

Notar que pese a la localía, también se utilizaron cantidades similares de tweets en los gráficos con respecto a los resultados. Es decir que se utilizaron cantidad similares de tweets en perdidos de local como de visitante.

9.2.2. ¿Podemos darnos cuenta en las horas previas al partido la dificultad de tal al disputarlo en condición de local? ¿Y en condición de visitante?

Nos surgió la pregunta si podemos darnos cuenta en las horas previas al partido la dificultad de tal. Es decir, los hinchas que tienen por delante un partido etiquetado como fácil ¿Tienen un mejor humor que los hinchas que tiene por delante un partido más exigente?

Nuestra hipótesis para esta pregunta es qué tanto de local como de visitante, los hinchas que tienen por delante un partido más fácil tienen un mejor humor que los hinchas con un partido difícil en las horas previas al partido, ya que tienen más posibilidad de ganar el encuentro a disputarse próximamente.

En Figure 9.5 y Figure 9.6 se puede ver cómo se cumplen nuestra hipótesis tanto de local como de visitante utilizando la métrica MeaningCloud normalizada.

El pvalue de comparar los valores de la métrica meaning cloud normalizada de todos los partidos etiquetados como fáciles contra los partidos etiquetados como difíciles en condición de local es 1.063e−7.

El pvalue de comparar los valores de la métrica MeaningCloud normalizada de todos

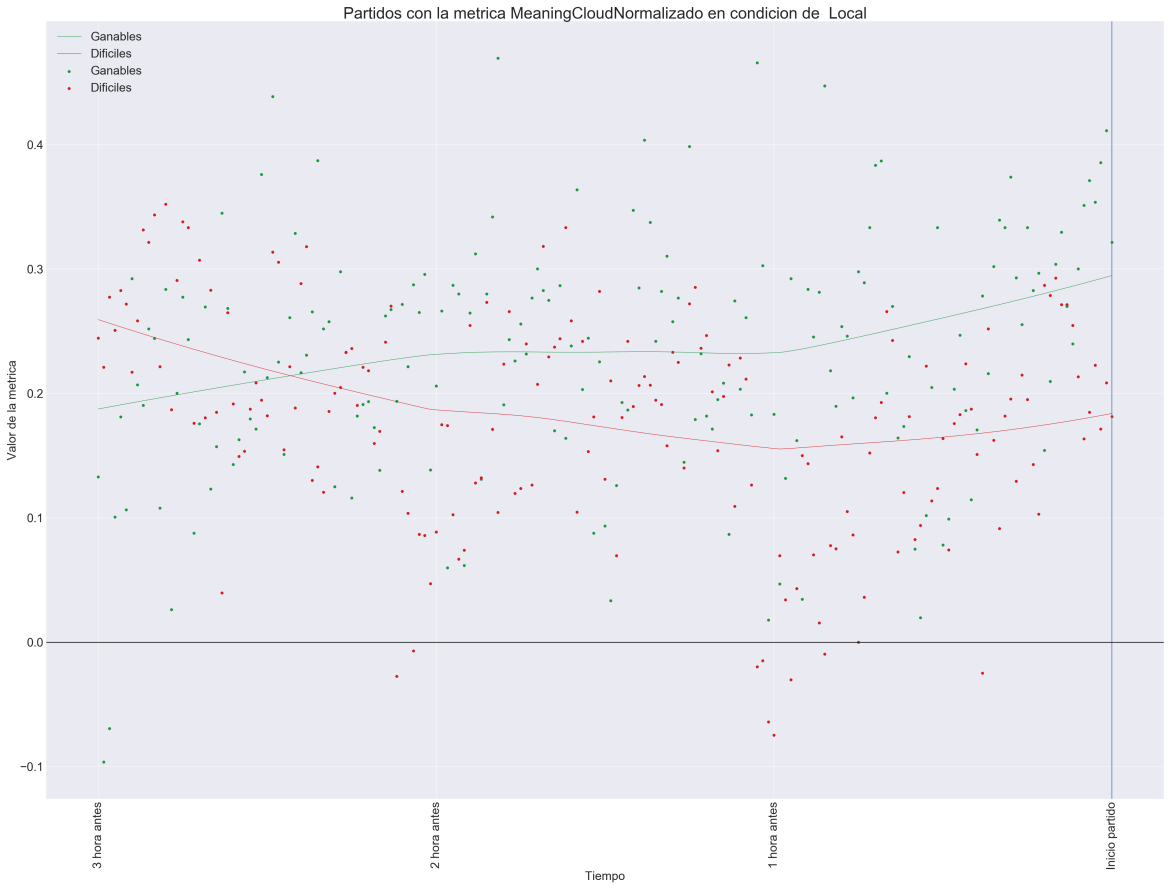

Fig. 9.5: Minuto a minuto desde 3 horas antes del comienzo del partido hasta su comienzo filtrado por dificultad del partido en condición de local

	Ganables	Dificiles
Media	0.217	0.178

Tab. 9.9: Media comparando por dificultad utilizando la métrica MeaningCloud normalizada en condiccion de visitante en la previa

Fig. 9.6: Minuto a minuto desde 3 horas antes del comienzo del partido hasta su comienzo filtrado por dificultad del partido en condición de Visitante

	Ganables	Dificiles
Media	0.231	0.155

Tab. 9.10: Media comparando por dificultad utilizando la métrica MeaningCloud normalizada en condiccion de local en el post partido

los partidos etiquetados como fáciles contra los partidos etiquetados como difíciles en condición de visitante es 0.0002. En Figure 9.5 que miramos en condición de local, utilizamos en total 159.236 tweets que se dividen 43.466 en partidos fáciles y 115.770 en encuentros etiquetados como difíciles. En Figure 9.6 que miramos en condición de visitante, utilizamos en total 135.586 tweets que se dividen 24.452 en partidos fáciles y 111.134 en encuentros etiquetados como difíciles.

Notar que media hora antes del comienzo del partido en condición de visitante, podemos ver que no se cumple nuestra hipótesis. Nos pareció un fenómeno extraño a destacar que no le encontramos una explicación.

9.2.3. ¿Podemos darnos cuenta en las horas siguientes al partido la dificultad de tal al disputarlo en condición de local? ¿Y en condición de visitante?

Por otro lado nos surgió una pregunta similar, la cual es si podemos darnos cuenta en las horas siguientes al partido la dificultad de tal. Es decir, ¿los hinchas que tuvieron un partido etiquetado como fácil tienen un mejor humor que los hinchas que tuvieron un partido más exigente luego de la finalización del encuentro?

Nuestra hipótesis para esta pregunta es similar a la anterior, es decir qué tanto de local como de visitante, los hinchas que tuvieron un partido más fácil tienen un mejor humor que los hinchas con un partido difícil en las horas siguientes al partido por la misma razón que en las horas previas.

En el gráfico Figure 9.7 se puede ver cómo se cumplen nuestra hipótesis de local utilizando la métrica de MeaningCloud normalizada.

El pvalue de comparar los valores de la métrica MeaningCloud Normalizada de todos los partidos etiquetados como fáciles contra los partidos etiquetados como difíciles en condición de local es $5.871e{-}22$.

En Figure 9.7 que miramos en condición de local, utilizamos en total 254.149 tweets que se dividen 64.547 en partidos fáciles y 189.602 en encuentros etiquetados como difíciles.

Sin embargo, podemos ver en Figure 9.8 que no se cumple en condición de visitante ya que se cumple tan solo en la primera media hora.

El pvalue de comparar los valores de la métrica MeaningCloud Normalizado de todos los partidos etiquetados como fáciles contra los partidos etiquetados como difíciles en condición de Visitante es 0.015.

En Figure 9.8 que miramos en condición de visitante, utilizamos en total 248.749

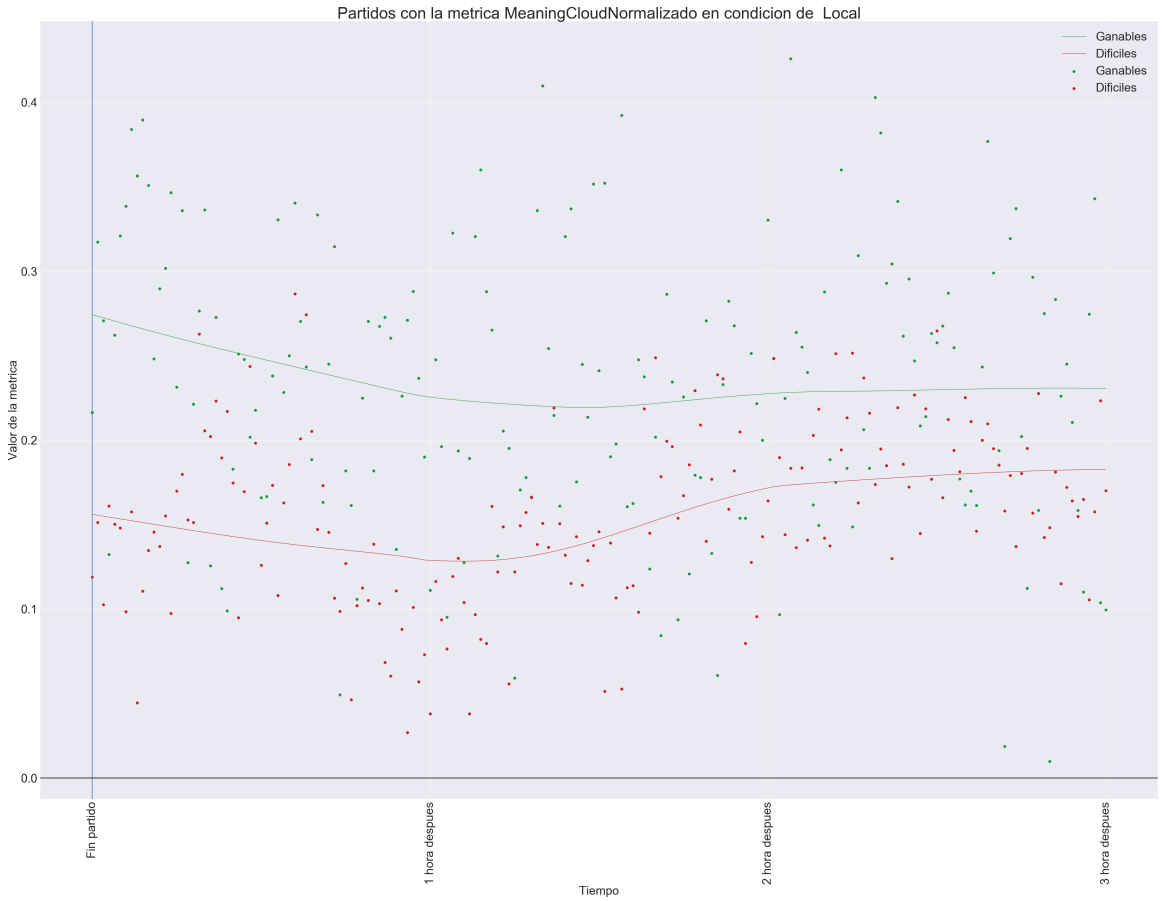

Fig. 9.7: Minuto a minuto desde la finalizacion del partido hasta 3 horas después de tal filtrado por dificultad del partido en condición de local

	Ganables	Dificiles
Media	0.114	0.137

Tab. 9.11: Media comparando por dificultad utilizando la métrica MeaningCloud normalizada en condiccion de visitante en el post partido

Fig. 9.8: Minuto a minuto desde la finalizacion del partido hasta 3 horas después de tal filtrado por dificultad del partido en condición de visitante

	Ganables	Dificiles
Media	0.02	0.026

Tab. 9.12: Media comparando por dificultad utilizando la métrica Facebook en condiccion de visitante en el post partido de partidos ganados

	Ganables	Dificiles
Media	0.259	0.342

Tab. 9.13: Media comparando por dificultad utilizando la metrica Meanincloud Normalizada en condiccion de visitante en el post partido de partidos ganados

tweets que se dividen 57.717 en partidos fáciles y 191.032 en encuentros etiquetados como difíciles.

Pensamos que lo sucedido en visitante se debía a la desproporción en cantidad de tweets, pero en ambas condiciones hay el triple de tweets en partidos etiquetados como difíciles.

Para tratar de entender lo sucedido en condición de visitante, vimos la cantidad de partidos y tweets por cada resultado en ambas dificultades. En condición de visitante, la cantidad de partidos etiquetados como ganables que se ganaron son 11 generando 26.703 tweets, se empataron 7 generando 16.315 tweets, y se perdieron 2 generando 14.699 tweets. En la misma condición pero partidos etiquetados como difíciles, se disputaron 9 que terminaron en victoria generando 78.972 tweets, 16 derrotas generando 95.546 tweets y 4 pardas utilizando 16.514 tweets. Pese a tener estos datos, no encontramos una posible explicación.

9.2.4. ¿Se festeja más ganar un partido etiquetado como difícil o fácil en condición de local? ¿Y en condición de visitante?

Por último nos preguntamos ¿qué se festeja más? ¿Ganar un partido etiquetado como difícil o fácil? Nuestra hipótesis es que se festejan más los partidos etiquetados como difíciles ya que genera mayor satisfacción en el hincha, ya sea de local como de visitante.

Utilizamos para los gráficos con encuentros ganados en condición de visitante, 128.425 y 41.498 tweets correspondientes a partidos difíciles y fáciles respectivamente.

Se puede ver en Figure 9.9 donde se muestran los partidos en condición de visitante que se ganaron como indica nuestra hipótesis.

El pvalue de comparar los valores de la métrica Facebook luego de la finalización del encuentro durante 8 horas utilizando partidos etiquetados como dificiles y faciles que terminaron con una victoria en condición de visitante es $2.190e{-}15$.

El pvalue de comparar los valores de la métrica Meaning Cloud Normalizado luego

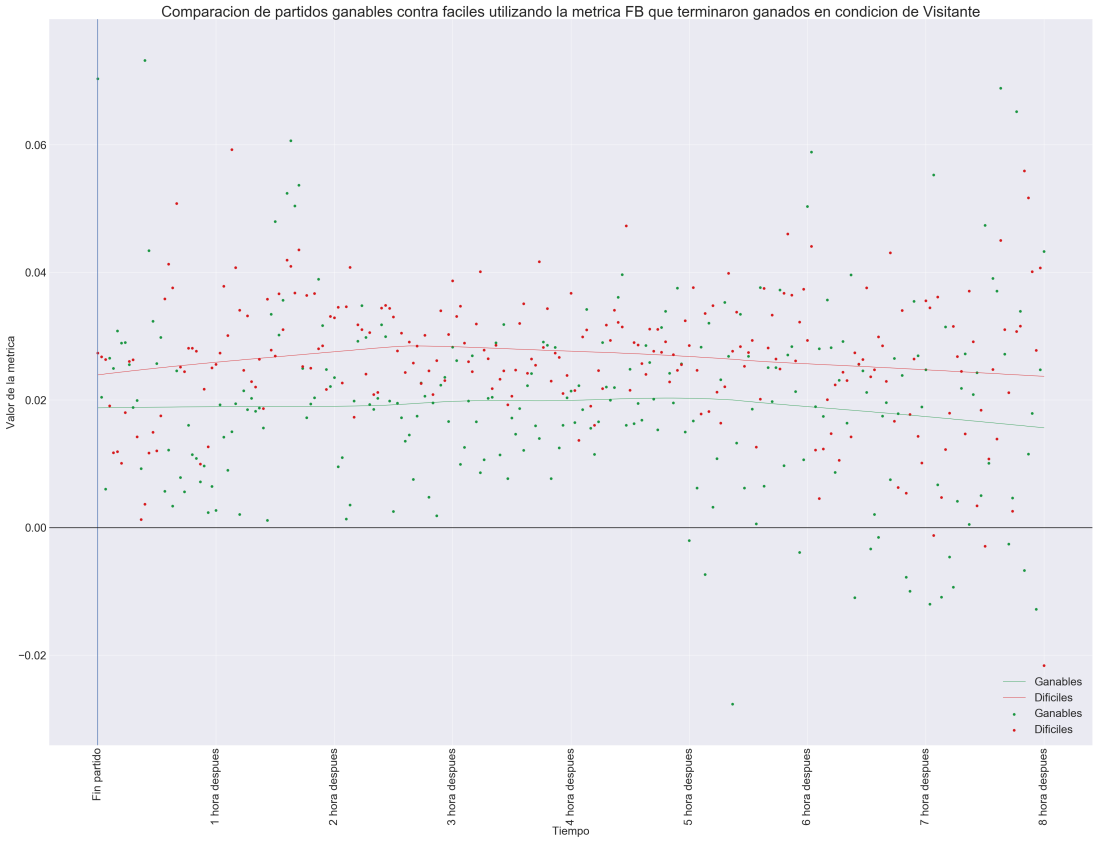

Fig. 9.9: Minuto a minuto de la métrica Facebook desde la finalizacion del partido hasta 8 horas después de tal filtrado por dificultad del partido en condición de visitante utilizando solo partidos que terminaron en victoria

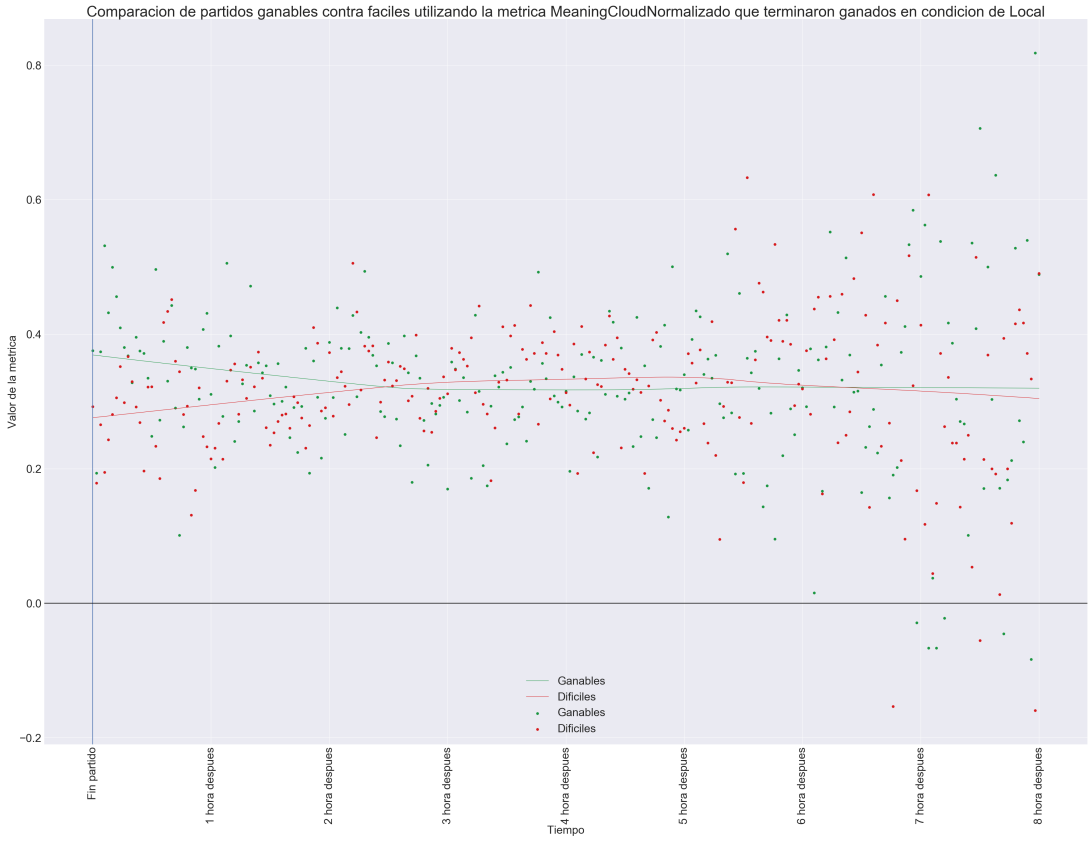

Fig. 9.10: Minuto a minuto de la metrica MeaningCloud normalizada desde la finalizacion del partido hasta 8 horas después de tal filtrado por dificultad del partido en condición de Local utilizando solo partidos que terminaron en victoria

de la finalización del encuentro durante 8 horas utilizando partidos etiquetados como dificiles y faciles que terminaron con una victoria en condición de visitante es 4.204e−15.

Sin embargo, no se puede apreciar lo mismo en condición de local.

Para Figure 9.10 en esta condición utilizamos 166.415 y 74.130 tweets correspondientes a partidos dificiles y faciles respectivamente, manteniendo que hay más del doble de cantidad de tweets en partidos difíciles.

El pvalue de comparar los valores de la métrica MeaningCloud normalizada luego de la finalización del encuentro durante 8 horas utilizando partidos etiquetados como difíciles y fáciles que terminaron con una victoria en condición de local es 0.242 lo que nos afirma lo viste en el gráfico: las muestran no son significativamente distintas a tal punto que las medias son muy similares.

No encontramos una explicación de porqué sucede esto.

	Ganables	Dificiles
Media	0.322	0.315

Tab. 9.14: Media comparando por dificultad utilizando la metrica Meanincloud Normalizada en condiccion de local en el post partido de partidos ganados

	Ganables	Difíciles
Media	0.09	0.076

Tab. 9.15: Media comparando por dificultad utilizando la métrica Meanincloud Normalizada en condiccion de local en el post partido de partidos perdidos

9.2.5. ¿Deprime más perder un partido etiquetado como difícil o fácil en condición de local? ¿Y en condición de visitante?

También nos preguntamos ¿que deprime más? ¿Perder un partido etiquetado como complicado o sencillo? Nuestra hipótesis es que deprime mas perder un partido etiquetado como sencillo ya que genera bronca sin importar donde se disputa el encuentro. Mostramos solamente el gráfico en condición de local ya que son muy similares a los de condición de visitante.

El pvalue de comparar los valores de la métrica MeaningCloud normalizada luego de la finalización del encuentro durante 8 horas utilizando partidos etiquetados como difíciles y fáciles que terminaron con una derrota en condición de local es 0.351 lo que nos afirma lo viste en el gráfico: las muestran no son significativamente distintas a tal punto que las medias son muy similares.

Utilizamos para Figure 9.11 con encuentros perdidos en condición de visitante, 149.231 y 19.658 tweets correspondientes a partidos difíciles y fáciles respectivamente. Mientras que en condición de local, 115.863 y 9.928 tweets con la misma correspondencia. Notar que hay mucha más cantidad de tweets pertenecientes a partidos difíciles que sencillos. En los gráficos anteriormente mencionados no podemos encontrar una respuesta a que deprime más en ambos tipos de localías, validado con el test. Notar que la varianza de los valores de las métrica es mayor en los partidos etiquetados fáciles. Pensamos que tal variación corresponde a que hay poca cantidad de tweets produciendo que no se logre llegar al promedio estable en cada minuto y por ende, no se llegó a formar una linea mas representativa como para poder comparar con los partidos con mayor dificultad.

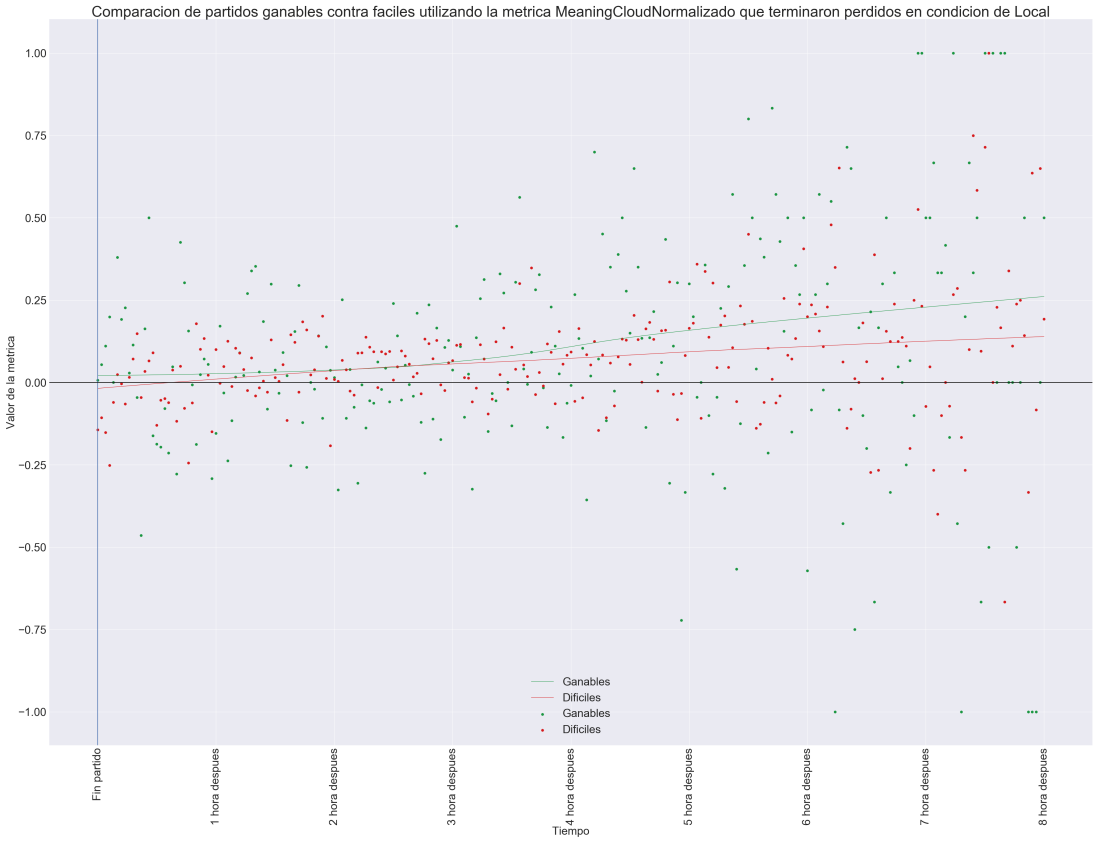

Fig. 9.11: Minuto a minuto de la metrica MeaningCloud normalizada desde la finalizacion del partido hasta 8 horas después de tal filtrado por dificultad del partido en condición de Local utilizando solo partidos que terminaron en derrotas

10. ANEXO DE DESARROLLO

10.1. Métodos fallidos

Para conformar el ground truth, primero realizamos una aplicación web donde pedimos a amigos, conocidos, y familiares que colaboren con el etiquetado. Lamentablemente, en esta opción invertimos mucho tiempo que no nos resultó útil.

Además de la página web, probamos realizar un perfil de Twitter que le consulta a seguidores de las cuentas semillas de que equipo son. Mediante un algoritmo, generamos tweets diferentes que el bot publicaba en Twitter cada cierto tiempo aleatorio pero como los tweets incluían una mención a un usuario, y no siempre nos responden la consulta, Twitter detecta nuestro bot como Spam y da de baja la twitter app que utiliza el bot. Incluso, nos llegó a cerrar los perfiles de Twitter de los bots.

También probamos realizar una encuesta de Twitter donde tenes que hacer RT al tweet donde indica tu equipo pero no logramos muchos perfiles etiquetar así ya que poca gente lo realizó. Lo bueno de esta forma es que sabíamos que una persona votaba de forma legítima ya que usaba su cuenta.

Incluso armamos una encuesta y lo compartimos para que la gente se auto-etiquete pero tampoco logramos etiquetar mucha gente de esta forma a pesar de que era insegura ya que cualquiera podía etiquetar incorrectamente a otra persona y no tenia ningun tipo de validación.

También le consultamos a los colaboradores que se registraron con Twitter en la página web de que equipo eran mediante Twitter pero, también, obtuvimos pocos ya que eran pocos los que usaron Twitter para jugar.

10.2. Herramienta para obtener información de Twitter

Para obtener información de Twitter, la red social brinda una API [1] con un límite para cada Twitter App, que es la forma de integrarse con tal API. Por un lado hay algunas restricciones que no se pueden solucionar mediante esta API, como por ejemplo la restricción que se pueden obtener hasta los 3200 tweets más recientes de cualquier perfil [2]. En caso de querer más tweets se puede utilizar servicios pagos como por ejemplo él brindado por GNIP [3]. Por otro lado hay algunos límites que impone la REST API que se pueden solucionar. Por ejemplo, la API permite 15 consultas por Twitter App para obtener a quien sigue un perfil en un lapso de 15 minutos [4] y capaz con una consulta no alcanza ya que cada consulta devuelve de a 5000 perfiles. Por ende, si se quiere obtener de 1000 perfiles a quienes siguen, se necesita al menos 67 ventanas de 15 minutos, es

[1] API de Twitter
[2] Método para obtener los tweets de un usuario
[3] GNIP
[4] Método para obtener a quien sigue un usuario

decir 17 horas aproximadamente,suponiendo que todos los perfiles siguen a menos de 5000 usuarios y es mucho tiempo para una poca cantidad de perfiles. La solución que encontramos fue crear muchas Twitter Apps y administrar cuantas consultas disponibles tiene cada una mediante un pool de conexiones.

Esta herramienta tiene la capacidad no solo de obtener tweets por primera vez de un usuario, sino también de conseguir nuevos tweets de un usuario que ya descargo tweets manteniendo los tweets ya descargados. Esto último resulta muy útil para no volver a descargar tweets ya descargados ya que se utilizó esta herramienta en distintas fechas y además con perfiles con muchos tweets, de esta forma se puede conseguir más de los 3.200 tweets permitidos por la api de Twitter. Por ejemplo, supongamos que para un usuario el 1ero de Enero se descargaron sus últimos 3.200 tweets, y al intentar el 1ero de Marzo el perfil tiene 10.000 tweets nuevos desde el último tweet descargado. No se va a poder descargar los 10.000 nuevos tweets por la limitación de la api de twitter, pero se va a poder descargar los últimos 3.200 tweets distintos a los primeros 3.200 logrando obtener 6.400 tweets para este usuario.

Como necesitamos poder hacer consultas rápido sobre los tweets, levantamos una base de datos utilizando Mongo. Por ejemplo, necesitamos obtener todos los tweets que pertenecen a un equipo que se realizaron durante un partido de tal equipo para realizarles análisis de sentimiento a los tweets. Optamos más que nada por Mongo ya que finalmente tenemos 26.601.818 Tweets y al tener tener ese volumen que es no estructurado, se recomienda utilizar bases no relacionales.

10.3. Herramienta para analizar el sentimiento de los tweets

Luego de tener nuestro ground truth de perfiles de Twitter etiquetados con su respectivo club por cual hincha, y tener los tweets de estos hinchas, necesitamos tener el sentimiento asociado a cada tweet para finalmente poder analizar el comportamiento de los hinchas durante los partidos. Como mencionamos en la sección de Análisis de sentimiento de Tweets en el estado del arte, utilizamos el servicio de Meaning Cloud. Ya que es un servicio online, nos comunicamos con tal mediante consultas HTTP [5] donde nos resultó clave el parámetro del nombre del modelo de análisis de sentimiento a usar para poder adaptar el producto a nuestro contexto ya que sino etiquetaba incorrectamente tweets sencillos como el festejo de un gol.

La respuesta de la consulta esta en formato JSON [6], similar a la respuesta de Twitter al consultar por tweets. Nosotros almacenamos esta respuesta en disco pero además cargamos el valor de sentimiento que responde la consulta en el registro del tweet en nuestra base de datos para nuevamente poder obtener más rápido los sentimientos asociados a un conjunto de tweets que cumplen ciertas características, por ejemplo, estar entre dos fechas y su escritor pertenece a un club en particular.

La respuesta del servicio nos devuelve un par de resultados pero utilizamos los siguientes:

[5] Parametros de la solicitud de MeaningCloud
[6] La respuesta de MeaningCloud

- Score_tag, que indica la polaridad del tweet. Los posibles resultados son muy positivo (P+), positivo (P), neutro (NEU), sin sentimiento (NONE), negativo (N), muy negativo (N+).
- Polarity_term_list, que es una lista de palabras con el score_tag que se le encontró en la frase ya que la misma palabra puede tener distintos valores de score_tag. Por ejemplo, no es el mismo score_tag que se la asigna a la palabra crisis en la frase 'La crisis global está terminando', ya que crisis tiene un significado positivo mientras que en la frase 'Hay una gran crisis', la connotación de la palabra crisis es negativo.

Bibliografía

[1] Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's @wally?: A classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 11–20, New York, NY, USA, 2013. ACM.

[2] Valentina Beretta, Daniele Maccagnola, Timothy Cribbin, and Enza Messina. An interactive method for inferring demographic attributes in twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 113–122, New York, NY, USA, 2015. ACM.

[3] Puneet Singh Ludu. Inferring latent attributes of an indian twitter user using celebrities and class influencers. In *Proceedings of the 1st ACM Workshop on Social Media World Sensors*, SIdEWayS '15, pages 9–15, New York, NY, USA, 2015. ACM.

[4] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

[5] Todd Bodnar and Marcel Salathé. Validating models for disease detection using twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 699–702, New York, NY, USA, 2013. ACM.

[6] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA, 2012. ACM.

[7] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 583–592, New York, NY, USA, 2015. ACM.

[8] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 841–842, New York, NY, USA, 2010. ACM.

[9] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.

[10] Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A. Smith. Predicting the nfl using twitter. In *Proc. ECML/PKDD Workshop on Machine Learning and Data Mining for Sports Analytics*, 10 2013.

[11] Stylianos Kampakis and Andreas Adamides. Using twitter to predict football outcomes. 11 2014.

[12] T. Yang, D. Lee, and S. Yan. Steeler nation, 12th man, and boo birds: Classifying twitter user interests using time series. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 684–691, Aug 2013.

[13] Morgan Jeffrey Sloan Luke Burnap Pete, Housley William. Social media analysis, twitter and the london olympics 2012. volume 1, 03 2014.

[14] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. 2010.

[15] Sitaram Asur and Bernardo Huberman. Predicting the future with social media. In *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, volume 1, 03 2010.

[16] Naushad UzZaman, Roi Blanco, and Michael Matthews. Twitterpaul: Extracting and aggregating twitter predictions. 11 2012.